



Data mining techniques in higher education research : The example of student retention.

BURLEY, Keith Martin.

Available from the Sheffield Hallam University Research Archive (SHURA) at:

<http://shura.shu.ac.uk/19412/>

A Sheffield Hallam University thesis

This thesis is protected by copyright which belongs to the author.

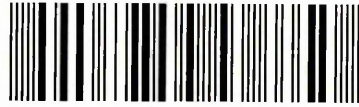
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Please visit <http://shura.shu.ac.uk/19412/> and <http://shura.shu.ac.uk/information.html> for further details about copyright and re-use permissions.

Adsetts Centre City Campus
Sheffield S1 1WB

101 857 339 9



REFERENCE

ProQuest Number: 10694293

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10694293

Published by ProQuest LLC (2017). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Data Mining Techniques in Higher Education Research The Example of Student Retention

Keith Martin Burley

**A dissertation submitted in partial fulfilment of the requirements of
Sheffield Hallam University
for the degree of Doctor of Education**

December 2006

Abstract

"Data mining is one of the faster growing fields in the computer industry. Once a small interest area within computer science, it has quickly expanded into a field of its own." (Westphal & Blaxton, 1998: 5)

Data Mining has been used for more than a decade in a variety of differing environments. It takes an inductive approach to data analysis in that it is concerned with the extraction of patterns from the data often without preconceived ideas. Data mining is part of the field of Business Intelligence, a subject area that the author is familiar with and has taught for many years. He believes that the application of data mining techniques has much to offer within the context of higher education. However, there is little evidence that these well established techniques have previously been applied to the sphere of higher education.

Student retention is a hot issue in higher education at the moment. It is for this reason that the author chose to establish the power of data mining techniques in higher education using the examination of student retention issues as a vehicle.

The field of student retention has been well documented over the years. Contemporary authors such as McGivney (1996), Moxley et al (2001), Yorke (1999) and Yorke & Longden (2004) have examined strategies and derived intervention techniques aimed at assisting students to adapt to university life. As the proportion of students entering Higher Education has increased there has been an increasing awareness that universities need to adapt to the changing profile of these students.

The data was collected via an online questionnaire administered to a large group of computing students at Sheffield Hallam University and similar institutions. The collected data was explored using Data Mining techniques including Decision Trees, Market Basket Analysis and Cluster Analysis.

This study sought to explore interrelationships between factors that contribute to student attrition and hence establish the demographics of at-risk students. The use of data mining techniques was found to be highly effective, having found most of the primary issues established in previous research. It went on to find the strongest relationships between them, corresponding well to findings from previous research using standard statistical techniques.

The author believes that he has established the power of data mining techniques in higher education and recommends further areas where it could be used profitably.

"Knowing how to exploit data effectively can help you to use available technologies to reveal the hidden patterns and trends contained therein."
(Westphal & Blaxton, 1998: xxi)

Contents

Abstract	ii
Acknowledgements	vi
Glossary	vii
Chapter 1: Introduction	1
Chapter 2: Focus of the Research	3
2.1 Aim	3
2.2 Objectives	3
2.3 Research Rationale	3
2.3.1 The Importance of the Work	3
2.3.2 Educational Importance	4
2.3.3 Financial Importance	6
2.3.4 The Interested Parties	6
2.3.5 Informing Professional Practice	7
2.3.6 Tools for Data Collection and Analysis	7
Chapter 3: Previous Research.....	8
3.1 Reviewing Location Research	8
3.2 Causal Problems	11
3.2.1 Widening Participation	11
3.2.2 Pressures to Improve Retention Rates	12
3.2.3 Financial Pressures	13
3.2.4 Part-time Work	13
3.2.5 More Flexible Study	14
3.2.6 Difficulties Measuring Attrition	14
3.2.7 Student Development	16
3.3 Modelling	17
3.3.1 Introduction	17
3.3.2 Tinto's Model	18
3.3.3 The Pathway to Retention Model	19
3.3.4 Isolating the Main Factors for Attrition	19
3.4 Intervention	22
3.4.1 Introduction	22
3.4.2 Improving Teaching & Learning	22
3.4.3 Supporting Learning & Supporting Students	22
3.4.4 The Beatty-Guenter Retention Strategy Model	23
3.4.5 Student Diversity	24
3.4.6 Footnote	25
3.5 Success and Barriers to Success	25
3.6 Student and Institutional Adaptation	27
3.7 Good Practice	28
Chapter 4: What is New in this Research.....	29
4.1 Introduction	29
4.2 What is Data Mining	30
4.2.1 The Concept	30
4.2.2 Data Mining as an Inductive Technique	30
4.2.3 Data Analysis Testing	31
4.2.4 Underlying Activities	33
4.2.5 Types of Learning	34
4.3 Data Mining in Industry	40
4.4 Data Mining in Higher Education	41
4.5 Summary	42
Chapter 5: Rationale of Approach.....	43
5.1 Rationale	43
5.2 Research Methods	43
5.3 Student Issues	44
5.4 Tutor Intervention	45

5.5	Recommendations	46
5.6	Ethical Issues	46
5.7	Educational Issues	47
5.8	General Problems	47
5.9	Research Overview	48
Chapter 6:	Initial Interviews	49
6.1	Introduction	49
6.2	Course	49
6.3	Institution	50
6.4	Academic	51
6.5	Personal	52
6.6	Motivational	53
6.7	Finance	54
Chapter 7:	Devising the Questionnaire	56
7.1	Introduction	56
7.2	Types of Question	56
7.2.1	Open-ended Questions	56
7.2.2	Closed-ended Questions	56
7.2.3	Choice of Question Type	57
7.3	Demographic Questions	57
7.4	Problem Questions	59
7.5	Administering the Questionnaire	61
7.6	Summary	61
Chapter 8:	A Methodology to Use	62
8.1	Introduction	62
8.2	CRISP-DM – A Methodology	63
8.2.1	History	63
8.2.2	Reference Model	64
8.2.3	The Stages of the Reference Model	64
Chapter 9:	Mining the Data	70
9.1	Business Understanding	70
9.2	Data Understanding	73
9.3	Data Preparation	79
9.4	Modelling	81
9.4.1	Rule 1	82
9.4.2	Rule 2	87
9.4.3	Rule 3	98
9.5	Evaluation	109
9.5.1	Introduction	109
9.5.2	What has been found	109
9.5.3	Summary	111
9.5.4	Vulnerable Group	112
9.6	Deployment	112
Chapter 10:	Focus Group Recommendations	113
10.1	Introduction	113
10.2	Discussion of Findings	113
10.3	Summary of Initial Finding	115
10.4	Follow-up	115
10.5	Recommendations	118
10.5.1	Problem Areas	118
10.5.2	Suggested Recommendations	118
Chapter 11:	Reflective Summary	120
11.1	Preamble	120
11.2	Focus of the Research	120
11.3	Previous Research	121
11.4	What is New in this Research	122
11.5	Rationale of Methodological Approaches	122

11.6	Initial Interviews	123
11.7	Devising the Questionnaire	123
11.8	A Methodology to Use	124
11.9	Mining the Data	124
11.10	Post-Enquiry Interviews	125
11.11	Focus-Group Recommendations	125
11.12	Postscript	125
Chapter 12:	Data Mining Recommendations	126
12.1	Introduction	126
12.1.1	Online Data Capture	126
12.1.2	Use of Data Mining	127
12.1.3	Required Skills Level	127
12.2	Further Work Involving Data Mining	127
12.3	Further Exploration of the Collected Data	128
12.4	Extending the Retention Study	128
12.4.1	Preamble	128
12.4.2	Extending the Study within Computing Students	129
12.4.3	Extending the Study beyond Computing Students	129
12.5	Data Mining Other Available Data	129
12.5.1	Students Records	129
12.5.2	HESA Data	130
12.6	Possible New Studies Involving Data Mining	130
Chapter 13:	Conclusions	131
13.1	A Note of Guidance	131
13.2	Summary	131
13.3	Use of Data Mining	132
13.4	Summary of Data Mining Recommendations	133
13.4.1	Introduction	133
13.4.2	Data Collection Methods	133
13.4.3	Data Analysis Methods	133
13.4.4	Further Exploration of the Collected Data	133
13.4.5	Extending the Study	133
13.4.6	Data Mining Other Available Data	133
13.4.7	Possible New Studies Involving Data Mining	134
13.5	Postscript	134
References	135
Bibliography	139
Appendix A:	Student Interviews	140
Appendix B:	Semi-Structured Interview Schedule	148
Appendix C:	Demographic Choices	149
Appendix D:	Student Non-Completion Problems.....	152
Appendix E:	Questionnaire	153
Appendix F:	Input Database Schema	155
Appendix G:	Variables and Types	156
Appendix H:	Conversion Criteria.....	157
Appendix I:	Data Mining Dataset	158
Appendix J:	Problem to Variable Lookup Table	159
Appendix K:	Post Enquiry Interviews	160
Appendix K.1	Introduction	160
Appendix K.2	The Interviews	160
Appendix K.3	Analysis of the Interviews	161
Appendix K.3.1	Initial Impressions	161
Appendix K.3.2	A Closer Look	162
Appendix K.3.3	Looking at the Major Issues	163
Appendix K.4	Summary	163
Appendix L:	Structured Student Interviews	165
Appendix M:	Analysis of Post-Survey Student Interviews	167

Acknowledgements

I would like to dedicate this work to my wife Carolyn, whose tireless patience, endurance and encouragement over the last few years has allowed me to continue when at times completion seemed so far away. Also thanks to my three children, Rich, Kate and James for their constant encouragement.

My supervisors Prof Peter Ashworth and Prof Ranald Macdonald have been an inspiration to me. Without their gentle encouragement and wealth of research experience I doubt that I would have had the confidence needed.

I would also like to acknowledge the administrative staff of Sheffield Hallam University Faculty of Arts, Computing, Engineering and Sciences for their invaluable assistance in the retrieval of student data, so vital to the research. I signal out Jane Thompson for special mention. She put aside the pressures of her own work on many occasions to satisfy my ever increasing demands for information.

Thanks also go to a colleague Dharmendra Shadija for the assistance he gave with the development of the web pages.

I would also like to thank the student themselves, both for taking time to complete the questionnaire and agreeing to be interviewed. Many of them have shown a great deal of interest in my work asking frequently how it was progressing.

No acknowledgement is complete without speaking of my colleagues at SHU. in particular Mike Rimmington for sharing some of my academic load at times when going got tough, Fran Slack for her patience and Matthew Love for his enthusiasm and energy.

Glossary

ACES	The name given the SHU Faculty of Arts, Computing, Engineering and Sciences
Apriori	A Data Mining algorithm used in Market Basket Analysis
ASP	Microsoft's Web programming language
Attrition	The loss of students prior to the completion of their course
CEM	College of Emergency Medicine
CHURN	A term derived from 'change' and 'turn' meaning a customer moving to an alternative supplier
Clementine	Data Mining software produced by SPSS
Cluster	The grouping together of entities by common factors
Confidence	A measurement of certainty in Data Mining
CRISP-DM	<u>C</u> ross <u>I</u> ndustry <u>S</u> tandard <u>P</u> rocess for <u>D</u> ata <u>M</u> ining
CSV	Comma Separated Variable
Data Mining	A method of exploring data inductively
Deductive Learning	The evidence provided in the data is regarded as true
Dream Weaver	Web programming software produced by Macromedia
Dropout	Students leaving university early without completion of their course
EDA	Exploratory Data Analysis
Enterprise Miner	Data Mining software produced by SAS
ETL	The <u>E</u> xtract <u>T</u> ransform and <u>L</u> oad process used to clean data
Failure	Those students who are lost to HE and are unlikely to return in the foreseeable future
FARG	Faculty of ACES Retention Group
FE	Further Education
GRI	Generalised Rule Induction, a Data Mining algorithm used in Market Basket Analysis
HE	Higher Education
HEFCE	Higher Education Funding Council for England
HEI	Higher Education Institute
HESA	Higher Education Statistics Agency
HND	Higher National Diploma
Inductive Learning	Is concerned with the extraction of patterns from the data
ITPA	Information Technology Programme Area
KD	Knowledge Discovery
KDD	Knowledge Discovery in Databases
K-Mean	A Data Mining Clustering Algorithm
Kohonen Networks	A Data Mining Clustering Algorithm
MBA	Market Basket Analysis
MIS	Management Information System
NCR	National Cash Registers
Neural networks	A Data Mining Algorithm that tries to mimic the human brain
Retention	The act of maintaining students on their courses
Rule Association	A Data Mining algorithm used in Market Basket Analysis
SAS	A statistical software company
SEMMA	<u>S</u> ample, <u>E</u> xplore, <u>M</u> odify, <u>M</u> odel, <u>A</u> ssess
SHU	Sheffield Hallam University
SIG	The CRISP-DM Special Interest Group
SPSS	A statistical software company
SQL	Standard Queries Language for database interrogation
SQL-Server	Microsoft's Relational Database
Supervised learning	A Data Mining form of learning that involves the use of a given output
Support	In Data Mining, the amount of evidence found in the data to support the rule
Teradata	The Data Warehousing software produced by NCR
THES	Times Higher Education Supplement
UCAS	Universities and Colleges Admissions Service
Unsupervised learning	A Data Mining form of learning that does not involve the use of a given output
USA	United States of America

Chapter 1: Introduction

"Access to higher education is not only a matter of getting in to university; it is a matter of staying in and emerging in good standing." (HCSCEE, 2001)

Though this study is concerned with establishing issues that affect student retention it is the method of data analysis that is the main focus. The author seeks to establish data mining as a strong and valid method of inductive data analysis using the area of student retention as a vehicle.

Throughout this document the word 'retention' will be used when talking about maintaining students on their courses. However, the author is aware that this implies a deficit model. The word 'retention' is more managerially-oriented and its use can lead to losing sight of the student perspective. It focuses "on the effectiveness and efficiency of an institution or a system". (Yorke & Longden, 2004: 5) They go on to argue that the more students fund themselves and move towards 'life-long learning', the rationale for retention and completion as a performance indicator becomes weaker. A more appropriate word might be 'progression'. However, since the word 'retention' is in common usage and its specific meaning in a student context is well understood. It has therefore been decided to continue to use it in this study.

The word 'retention' means different things to different people, but in the context of students in higher education it is necessary to be clear about the meaning that is to be used in this study. Moxley et al (2001) believe that it is more than just addressing the academic requirements. That is, retention is keeping students and helping them to overcome the problems that they encounter on entering higher education and coming to grips with whom they are and where they are going. However educational standards must not be compromised in the process. Each "student must achieve academic requirements to persist and to graduate." (Moxley et al, 2001:5)

"The increasing rates of participation in higher education have a dramatic downside". (Moxley et al, 2001, cover)

The retention of students in higher education has been a highly contentious subject for many years (Martinez, 1996), and never more so now, in a climate of extending and widening participation. The non-completion rates have been steadily rising in Britain, but vary dramatically in the western world. Japan has the lowest non-completion rate at 10% with Britain double that at 20%. However, other countries fare much worse at 28% in Germany, 37% in the

United States and a massive 45% in France. Many British Universities and their counterparts throughout the western world are beginning to face the fact that action needs to be taken to stem this tide. It is in this atmosphere that the research was conceived. The following section sets out the tight boundaries within which this research was conducted.

The above figures should be taken with some degree of scepticism. Not all factors in each of the countries mentioned are equal. France, for instance takes a much larger percentage of students into higher education in the first year as does Italy. There they have a relatively open access to the first year of higher education. These students have to prove themselves before continuing to the subsequent years, hence the much higher non-completion or attrition rate. (Yorke, 1999) There is a similar picture in the United States. In Germany there is a different picture; students often take an extended period of time to gain their first degree and hence the figures are difficult to compute. (Yorke, 1999) We are therefore not equating like with like. In Japan there are both public and private universities; the above statistics only tell part of the story. There are also significant cultural differences from other western states. A more detailed look at the Japanese system might be worthy of investigation. Such an investigation was beyond the scope of this study.

Moore (1995) did a study based on Sheffield Hallam University. However, a number of the findings of that study are now outdated, not just because of the passage of time but since most of the recommendations made have already been implemented at the university.

It might be worthy of note that there needs to be an expectation of student personal responsibility and that these students are ready to be retained in higher education (HE) having mastered the role of a student. (Moxley et al, 2001) This implies a partnership of cooperation between the university and the student where the university accepts that they have a part to play in the process of increased progression and the student accepting that they have obligations and responsibilities within the process. Peelo et al (2002:123) believe that failure in HE should be accepted as a "normal and desirable part of the learning experience" and that tutors should not seek to protect students from it.

"A measure of retention indicates the proportion of students not successfully transferring from one stage of a course to the next." (HCSCEE, 2001)

Chapter 2: Focus of the Research

2.1 Aim

In the light of increased student participation in HE, an increasingly large proportion of Sheffield Hallam University (SHU) Information Technology Programme Area (ITPA) computing students are dropping out during or at the end of their first year. There is also a significant number in later years.

The aim of this study is to explore the issues that affect SHU ITPA computing students and contribute to dropping out, with the view to possible tutor intervention to improve retention rates. It is intended to test the value of using Data Mining techniques in HE using the vehicle of student retention.

2.2 Objectives

This aim now needs to be operationalised into a series of objectives. The achievement of these objectives and the methods by which these are achieved is the subject of later discussion. Below are six clear objectives that the study has sought to explore:

1. Identify from literature and secondary qualitative research, the student issues that contribute to dropping out;
2. Investigate quantitatively, by means of a questionnaire, how widespread these issues are;
3. Explore, by means of Data Mining techniques, the interrelationships between the issues;
4. Validate the results of the quantitative research within the ITPA by means of structured interviews;
5. Identify by the use of a ITPA tutor focus group, where tutor intervention can be used to help alleviate these issues;
6. In the light of the findings compile a list of recommendations on how the use of data mining techniques could be further developed.

These objectives fall into three main categories. The first four are student focused, the fifth is directed towards tutors and the final objective looks at how data mining techniques can be used to greater advantage in the future.

2.3 Research Rationale

2.3.1 The Importance of the Work

The work is important for a number of reasons, the principal ones being educational and financial. It is the author's belief that education is important in its own right; a mature, educated and informed society will be more open and less prejudiced. In addition there is a general perception, internationally, "that

economies are best served by maximising the level of education in the populace.” (Yorke, 1999:1)

Yorke (1999) believes that governments around the world are beginning to call higher education institutions (HEI) to account for the money that is invested and that they are being put under ever increasing pressure to keep public spending under control. In England the financial burden of higher education is gradually being moved from the government, through the local education authorities, to be the responsibility of the student. Successive governments have sought to change the funding structure, initially by the introduction of student loans in the mid 1990's and the introduction of a contribution to the tuitions fees introduced in 1998. Top-up fees, introduced in 2006, at the university's discretion with deferred payment collected out of taxation when a threshold income level is reached, adds a significant burden to the debt ridden student. The purpose of these measures is presumably to reduce the funding burden on government whilst still increasing the participation in higher education in England to their stated target of the fiftieth percentile.

2.3.2 Educational Importance

The research starts from the premise that student transition into higher education could be handled better, particularly in the changed climate of increased student participation. (Dearing, 1997) In many universities there is little provision for guidance of students in their first year to help them to adapt to the different environment that they find themselves in. In SHU such provision has been made but appears to be used inconsistently, depending on the individual tutors and their practices together with the take-up of services such as Educational Guidance. Most of the students are moving from the relatively protected environment of school or college, where they are guided through their educational experience and constantly monitored and encouraged if they fall behind or fail to attend. The atmosphere in university is much more relaxed and informal. Students who fail to attend lectures and tutorials are rarely identified until several weeks into the first semester (or term), by which time habits and bad practices have been formed which are difficult to change. Irretrievable damage might already have been done. In order to examine this we need to look beyond educational issues. (Moxley 2001) Many of the problems affecting the students are outside the educational process. “Most institutions recognise

that undergraduate education is much more than formal instruction and encompasses opportunities to develop socially, culturally, physically, spiritually and ethically.” (Moxley et al, 2001:58) This was also a major theme of the 1944 Education Act when compulsory secondary education for all was introduced.

Over the past decades the provision of higher education has gradually increased from a level of about 5% of the population in the early 1960’s to around 40% last year (2005). Following the 1997 Dearing Report the government has become committed to increase student participation in England to 50%. However, a significant number of students fail to complete their degrees (20% in Britain, see Chapter 1 above). At a national level, this could be looked upon as a failure to educate them to their full potential. At the university level it affects the completion rates, reduces funding and possibly future potential student numbers; at a student level there is the issue of personal failure and debt. Completion rates are one measure that may be used to compare one university against another when students are deciding on their university choices.

The growth of student admissions towards the government target of 50% intake into higher education appears to be holding up. The final figures for 2005-6 entry were up by 7.4% (27,825). In all a total of 405,369 applicants were accepted onto courses compared with 377,544 in 2004. (UCAS, 2006b) However the figures published by UCAS on 24th March 2006 show a 3.2% decrease in student applicants for 2006-7 entry compared with the figures at the same time the previous year. (UCAS, 2006a) It is yet to be seen whether the increase in tuition fees will have a real impact on student admissions. However, Computer Science is showing a different picture:

Year	Degree	HND
2004	11,659	1,627
2005	11,600	1,269
Reduction	-0.50%	-22.00%

(UCAS website)

As of 24th March 2006 the number of applicants for an undergraduate degree courses were down by 9.9%. HND applications were down by 30.8%. (UCAS, 2006b) However, there has been a move from HND to Foundation degrees over the last two to three years, so the HND figures can be explained. This decline is over all age groupings. Actual admissions for Computer Science still show a downward trend. The final enrolment figures for SHU Information

Technology Programme Area show a decline from 541 in 2005-6 to 436 in 2006-7. Whether this is a blip due to increased student fees is yet to be seen.

2.3.3 Financial Importance

For the university, every student that is lost is a reduction in income. Recruitment of students is an expensive operation and if a high proportion of these students drop out then this significantly affects the university's budgets. It might well be that the costs associated with efforts to retain students will be more than offset by the retention of students fees, perhaps yielding a healthy profit.

From a government (delegated to the local authority) perspective every student that drops out represents student financial support wasted. That is, monies that cannot be recovered from either the university or the student and which have not contributed to a successful student qualification. "Non-completion and delayed completion can be construed as inefficiencies in the use of public finances, and hence they become political issues." (Yorke, 1999:2) Yorke estimated that in the academic year 1994-5 (taking out returning students) the cost of non-completion was £91m which represented 3% of the total expenditure on higher education. (Yorke, 1999) No up-to-date figures have been found to date, but it is anticipated that the figure will be much greater.

From the student perspective debts have still accrued during their time in higher education. If they fail to complete they are left with these debts from student loans and bank overdrafts that still need to be repaid. Their capacity to repay is likely to be less than those who have completed their degrees as non-graduates' earning potential is lower.

2.3.4 The Interested Parties

The research population consists of computing students from SHU and some other similar universities. It focuses on the SHU computing students in ITPA within the Faculty of Arts, Computing, Engineering and Sciences (ACES).

The research will be of interest to the management of the SHU Faculty of ACES and other such institutions. An outcome of the research is a set of recommendations to be delivered to a focus group (See Chapter 10) for consideration. It is intended that these recommendations will form a basis of a discussion document in an effort to form a working practice document to be implemented within the faculty. It will, therefore be of general use to tutors and

support staff at the grass roots level. Moxley et al (2001) believe that it is important that the commitment to retention should be the responsibility of those who interact with students on a daily basis.

2.3.5 Informing Professional Practice

Since an outcome of the research is a set of recommendations, year tutors, module leaders and indeed individual lecturers may find them useful as a template. It is hoped that the template might also be of use to students. Involving tutors in the process in a real way will help to ensure that the recommendations will be implemented, at least in part, though implementation is not part of the study. Every effort will be made to encourage ownership of the recommendations by the tutors who helped to conceive and develop them.

Although the research is focused within the context of the ITPA in the Faculty of ACES within SHU it is expected that the findings may be used by others interested in the field from diverse British Universities. Though the research is primarily based in SHU it is believed that the findings should be relevant to a wider population, particularly from the 'New Universities'. It would, therefore, add a new dimension to the empirically-based knowledge of sociology in Higher Education. As a future extension to the study it should be possible to replicate the research to enhance transferability.

As stated in Chapter 1 the author has focused the research on the use of data mining as an analytical tool using student retention issues as a vehicle. Since data mining is a relatively new medium, there being very little documented research in the field of higher education, it is expected that the recommendations will be a significant starting point for further studies. The expected contribution to the body of knowledge will be the publication of the outcomes to a national and international academic audience. In this way it will add to the growing literature in the field and play its part in the review process of student pastoral and academic provision in higher education.

2.3.6 Tools for Data Collection and Analysis

The research falls into three main methodological areas:

- Qualitative analysis – from interviews
- Quantitative analysis of data – an online questionnaire (statistical analysis)
- Data Mining – an online questionnaire (exploring to find hidden patterns)

Chapter 3: Previous Research

3.1 Reviewing Location Research

There is a fair amount of research that has already been done in the field of Student Retention, much of which is in higher education. However the student perspective is changing, particularly in the light of three significant developments. Firstly, the introduction of students loans in the 1990's, secondly the introduction of student contributions to tuition fees introduced in 1998 (extended in 2006) and finally the ever increasing number of students entering higher education. It has been decided that any research conducted prior to the mid 1990's should be used with caution and any findings from such research should be backed up with more contemporary research.

At this stage three principal books are considered, Moxley et al (2001) entitled "Keeping Students in Higher Education", Yorke (1999) entitled "Leaving Early" and McGivney (1996) entitled "Staying or Leaving the Course". Each of these texts has some significant findings that will prove useful in this study.

Moxley et al (2001) conducted a largely qualitative study. It is written from the standpoint that "higher education should not be closed to those individuals who wanted to improve themselves" (Moxley et al, 2001: ix), a view with which this author concurs. It doesn't set out to offer a rigid framework for retention and recognises that there is a large variation in practice. In particular, it recognises that retention is about more than just achieving academic standards. It advocates that the students should recognise and come to terms with their place in the process. It sets out a "Pathway to Retention" outlining objectives to be achieved and supportive retention practices. This is covered in Section 3.3.3 below. Moxley et al (2001) believe that the HEI (higher education institutions) should offer proactive approaches to link retention and persistence to student development requiring individualisation for each student and that advisors and counsellors are essential to the process. (Moxley et al, 2001:91)

The research conducted by Yorke (1999) may be classified as using mixed methods. It uses a combination of qualitative and quantitative research techniques. He starts out by explaining that governments around the world are bringing higher education institutions to account for their expenditure on education. (Yorke, 1999:1) It looks at a number of different models defined to

tackle the subject of retention by such people as Tinto (1993), Napoli & Wortman (1997), Bean & Metzner (1985), Ozga & Sukhnandan (1998), Johnson & Buck (1995) and Long et al (1995). Its main research techniques were mail questionnaires and a telephone survey. It comes up with six principal issues that affect student dropout (Yorke, 1999:39-46):

1. Poor quality of student experiences

This is chiefly relating to quality of teaching received and the level of support given by both academic and other support staff.

2. Inability to cope with the demands of the programme

This centres around the stress related to the course of study due in part to a heavy workload and a lack of appropriate study skills often resulting in insufficient academic progress.

3. Unhappiness with the social environment

This might be caused by homesickness, accommodation problems or a dislike for the town or city in which the institution was located.

4. Wrong choice of programme

The programme of study might have turned out to be different from that envisaged and have little perceived relevance leading to disinterest.

5. Matters related to financial need

Financial needs and consequent term-time employment rated heavily here. This is often exacerbated by the lack of parental support.

6. Dissatisfaction with aspects of institutional provision

This covers the provision of a library, computers and specialist equipment. Those entering the institution through clearing might find things not as anticipated.

He appears to be committed to the expansion of higher education but offers a strong word of caution. "The opening up of higher education cannot be accomplished without risk of non-completion." (Yorke, 1999:110)

McGivney on the other hand conducted her research with mature students (those over 21 on commencement of study) in FE (Further Education) and HE (Higher Education) in 1995. Although this research is only just within the author's review remit, it does offer some insights into methodologies. It used a mixture of previous research, consultation with representatives from FE and HE, a postal survey, contact with a sample of Access Validating Agencies¹ and correspondence with other researchers in the field. She states that the 1990's brought a new climate and opportunities to the landscape of adult learners in HE, bringing about an "increasing flexibility in entry requirements." (McGivney, 1995:3) She defined the potential paths out of a course as:

¹ "Access" is a means of study that enables mature students to enter HE without the usual entry requirements

1. Non-starter

Those students who either fail to arrive at the start of the programme of study or drop-out within the first few weeks.

2. Informal withdrawal

This is perhaps the passive way to withdraw, by simply stopping attending. This might be gradual over a period of weeks or months. The student does not officially withdraw but simply fails to hand in coursework or attend examinations.

3. Transfer to other programmes

If a student finds the course of study inappropriate or too difficult then they might transfer to other programmes in that institution or a similar one. This can take place at any time during the course.

4. Academic failure

This category includes those students who have completed the course of study in a particular academic year but have failed to achieve the necessary standard to progress.

5. Formal withdrawal

At a particular stage in a programme of study a student might officially withdraw and either leave completely or defer. This is active withdrawal.

6. Non-continuer

This is where a student completes a particular year of study and then decides not to continue.

She then goes on to set out a strategy for retention.

All these studies come at the subject of retention from the premise that increased participation in higher education is a good thing. They offer a real insight into the reasons why students fail to complete and offer constructive advice on what might be done to assist these students and hence increase retention rates.

Extensive reading has revealed a number of recurring themes. They can be grouped into three major categories:

1. Causal Problems

The things that affect retention such as part-time work and other outside pressures;

2. Modelling

Attempts to create retention models such as that produced by Tinto;

3. Intervention

How support staff can interact with the students in an effort to increase retention.

In the following sections each of these are explored and factors associated with these are identified.

3.2 Causal Problems

3.2.1 Widening Participation

"The student population in higher education is becoming increasingly diverse as a result of widening participation, and this is set to continue with the Government target set at 50% ... by 2010." (Slack & Casey, 2002:1)

Most developed nations have placed much emphasis on the need to develop a "skilled, well-educated and flexible workforce." (Martinez, 1996:5) Martinez goes on to say that there is a consensus that Britain is falling behind many of our major economic competitors. These factors have led the government to set the above target. "Widening participation is central to economic prosperity and social cohesion." (Kennedy, 1997:15)

Widening Participation is a broad issue and is associated more with broadening the student intake to ensure that previously under-represented groups are attracted into higher education rather than just taking more students. "Participation must be widened and not simply increased." (Kennedy, 1997:15) Many authors have remarked that it is the post-1992 universities that are predominantly catering for these 'non-traditional' student entrants. (Archer, 2002) However, Archer argues that widening participation runs the risk of reproducing inequalities rather than tackling them. She goes on to say that working class students are likely to experience greater financial hardship and are more likely to take on term-time work. This is a theme taken up by Callender (2001). She remarks that although students from poorer backgrounds are not likely to contribute to tuition fees they are still more likely to end university with larger debts because of the lack of parental funding. She suggests that these groups are also more likely to be averse to debt and hence bow to the pressure to work longer hours during term-time.

It has been commented extensively that widening participation has increased the trend to stay living at home whilst studying. (Archer, 2002) This is particularly evident with Asian females (UCAS, 2002; Pugsley, 1998) and students from the lower socio-economic groups. (Connor et al, 2001) Forsyth and Furlong (2000) go on to say that going away to study is now looked upon as a necessity rather than a conscious choice, hinting that attending an HEI nearer to their home is preferable. Connor et al (2001), quoting HESA statistics, states that there was a 40% increase in home based students between 1997 and 1999, the point at which students fees were introduced.

It is clear then that there are two major consequences of widening participation; they are increased term-time employment and a greater number of 'home' students. Care must be taken here when looking at cause and effect, since widening participation comes with increased student fees. Thomas (2002) warns that greater diversity will lead to an increase in student withdrawal. However, it appears that this might not be the case. "HE has expanded but the proportional non-completion has remained relatively stable." (House of Commons, 2001: Para 1.11) Tight (1998) warns us of "victim blaming" and states that it is too easy to blame the "new students" for any future increase in early withdrawal. Home students, however may form very different relationships with the university and fellow students from those of non-local students. They may well have outside commitments and pressures that inhibit them from taking part in the extra-curricula activities of the traditional students. These students are often particularly hard to contact. (Slack & Casey, 2002)

3.2.2 Pressures to Improve Retention Rates

"In order to achieve government targets institutions will need to encourage wider participation, maintain standards and raise achievement rates." (House of Commons, 2001)

In the early 1990's the issue of student retention was given little importance. As the 90's continued however, there were increased pressures. There were strong moves to increase participation rates (widen participation). There were equally strong pressures to keep the students that were already enrolled (increase retention). However, Kennedy (1997) warns that this is providing more opportunities for those who have already achieved and helping them to continue to do so.

Improved retention is looked upon as being multifaceted, maintaining income for the colleges and universities, improving aspects of college life (Lalgree cited by Martinez, 1996) and providing success for students. There appears to be some progress. Some 77% of fulltime degree students were projected to achieve a degree at the institution in which they started which compares favourably with other countries. Because of these increased pressures and with the maintenance of educational standards in mind, the government has set targets to be achieved both for wider participation and achievement rates. (House of Commons, 2001)

3.2.3 Financial Pressures

"[It is] estimated the annual cost to the taxpayer of early student departure as around £100m". (Yorke, 1999:46)

Financial pressures are principally coming from two major sources. Firstly from the universities, as much of their continued funding is based on targets for wider participation and achievement rates. Secondly, with the introduction of student loans in the early 90's and tuition fees in 1997 (greatly increased in 2006) there is a significantly increased pressure on students to complete or to withdraw before debts become too great. This student pressure has been felt hardest by students from low-income groups where the parents are less able to fund a shortfall. (Thomas, 2002) It is also expected that the number of students applying to university may be significantly affected by the increase in fees and subsequent increase in debt. (Callender, 2003) This appears to be born out in reality with a reduction of 3.2% in overall applications to HE and a 9.9% reduction in applications to Computer related courses as of 24th March 2006. (See 2.3.2 above)

There has been much emphasis on hitting recruitment targets in order to maximise funding. However, Moore (1995) warns us to examine the pay-off between hitting the targets and recruiting the right students who have interest and commitment in their chosen programme of study. This may be just as true today as it was in 1995.

3.2.4 Part-time Work

It is not so long ago that term-time working by students was a breach of the rules. However it is now looked upon by many as a necessity in order to help fund study, therefore accepted as the norm. This is putting considerable pressure on a high proportion of students. The pressures on students to work at times when they have study commitments are often difficult to balance causing cries for more flexible study facilities. One such pressure is towards distance learning and the increased use of e-learning. (McInnis et al, 1995) Many universities are looking at ways that this can be best tackled.

However, the experience of part-time work is not necessarily all negative. As well as easing financial pressures it can provide excellent work experience. "It can give you a taste of different working environments, and provide a competitive edge for when you leave university and enter the workplace." (Aim Higher, 2006)

The universities can make more use of part-time working by integrating such work either formally within the curriculum, work placement (on a day a week basis or block) or less formally with such things as special initiatives and job shops with local employers and organisations. (Aim Higher, 2006)

Some universities offer schemes for part-time work within the university itself. Southampton University is one such institution as indeed is SHU. The schemes have been deliberately designed to accommodate student timetables. Students should not take on too much additional work though, as study time is very important. (University of Southampton, 2006)

3.2.5 More Flexible Study

One solution that has been suggested by a number of authors is that of an extended period of study. This has been a norm in Germany for a number of years. (Yorke, 1999) Kennedy (1997:23) talks about a "self-perpetuating society". This might mean breaks in the student education to allow them to top up their income, part-time study, distance learning or a combination of all of them. However, Moore (1995:40) warns that "[g]reater flexibility of study and widening student choice correlate with the need for more student guidance as do higher numbers of non-traditional entrants". This comes at a time when student-staff ratios are increasing.

3.2.6 Difficulties Measuring Attrition

Tait (2004:97) considers retention at two levels, the "single course or persistence through a programme of study". For the purpose of this study a student will be deemed to have failed to be retained if they do not complete a particular year of study and do not transfer to another course and/or institution. Attrition is deemed to be a measure of this non-completion. If we take this definition then we are faced with the problem of when precisely a student is deemed to have been lost. For instance, if he or she transfers to another course within the university are they lost? If they return to HE within a couple of years are they then considered to be found again? Universities have no real way of knowing whether their students are moving on to other universities or colleges when they leave as leaving is not always an active decision but more of a default due to circumstances. They simply stop attending. (See section 3.1, p10).

Ashby (2004:66) examines this dilemma and attempts to put a boundary on it. She suggests that viewing retention as “a measure of the percentage of students who gain a course credit or an award” is too narrow a definition. She goes on to offer three dimensions to retention, institutional, student and employer.

a) The Institutional Dimension

This refers to indicators an institution may use to measure retention and how well it is performing such as course completion rates. Influences such as government funding and the Quality Assurance Agency external reviews are crucial here as are university league tables, however crude they might appear to be. Because the government has put a huge investment into higher education and because high dropout rates are generally seen to be associated with poor institutional performance it is vital that institutions “develop their own internal measures to help them understand and interpret retention in the context of their mission and their student population.” (Ashby, 2004:66)

b) The Student Dimension

Course completion rates are important measures; they are used as a measure of student success but don't necessarily measure satisfaction. Recommendations from a Task Group (HEFCE¹ 02/15), places emphasis on the importance of student feedback on their learning experience including student satisfaction. The results should then be shared externally to allow prospective students informed choice.

Coldeway (1986) suggested a student success matrix that takes into consideration course completion and satisfaction with the course. In this way it is possible to distinguish between courses with low retention rates and low student satisfaction rates and courses with low retention rates and high student satisfaction rates. (Ashby, 2004:67) Moxley et al (2001) argues that retention is not just completing the course but more to do with becoming a successful learner, able to meet study goals. This plainly suggests that the major focus of student retention is to actively seek to assist the students to set their own study goals and achieve them, not just to stay the course.

¹ Higher Education Funding Council for England

c) The Employer Dimension

The employer dimension has become more important recently as the government looks for value for money. "Key skills that will contribute to employability are an important element in courses/awards offered by institutions." (Ashby, 2004:67) The introduction of Foundation Degrees and increased work placement make this much more relevant. This suggests a greater emphasis on the relevance of courses and a possible positive effect upon student retention.

3.2.7 Student Development

Much has changed in pre-higher education over the last couple of decades. Approaches to teaching and learning, assessment methods and subject material have all changed radically, particularly since GCSE replaced the O' level system and the introduction of modular A' levels. However according to Cloonan (2004:176) "recent research has shown how little has changed in universities ... as pre-existing hierarchies continue to thrive". Whether or not this is true is not part of this study.

Chickering (1969) came up with Seven Vectors of Student Development. He calls them vectors because each seems to have direction and magnitude. These were revisited by McEwen et al (1996) and two extra ones were added.

1. Developing Competence

This is the ability to cope and to achieve. This isn't just intellectual competence but physical and interpersonal competence as well.

2. Managing Emotions

As students move out of adolescence to adulthood they develop passion and commitment. They also learn to manage such things as anger and sexual desire.

3. Becoming Autonomous

Students develop independence, being able to take care of themselves both emotionally and practically as they "spring the nest"¹.

4. Establishing Identity

Who am I? The student gains a sense of self. This might have added poignancy for women and ethnic minorities where their roles in their ethnic society differ from those in the institution.

¹ 'Spring the nest' is a colloquial term that means leave the family home

5. Freeing Interpersonal Relationships

Acquiring tolerance and respect for others. Moving from dependency through a period of negotiation and understanding the need for inter-dependency and the mutual benefit of relationships.

6. Clarifying Purpose

Where am I going to be? Understanding career and life goals and making appropriate choices to achieve them.

7. Developing Integrity

Developing the ability to live with uncertainties and adapting to society norms and rules; developing a personally valid set of beliefs.

8. Interacting with the dominant culture

We are what we buy/consume, as the market-based culture threatens to turn us into mere consumers.

9. Developing spirituality

Experiencing ourselves as spiritual beings; being in touch with our spiritual centres and possessing an inner peace.

These issues have a significant bearing upon retention, in that they recognise a gradual transition from adolescence to adulthood; in particular, the acceptance by students of responsibility for their own education. Indeed an 'Extended Induction' scheme devised to ease student transfer into higher education is to be trialled in the coming academic year (2006-7) at SHU within the Faculty of ACES. This will involve approximately 300 students during the first eight weeks of semester 1.

3.3 Modelling

3.3.1 Introduction

Over the years there have been a number of attempts to model the student attrition path. Zepke & Leach (2005), state that the authors of the different models can be roughly divided into two camps. Firstly there are those who seek to revise and enhance Tinto's model¹. (Cabrera et al, 1992; Braxton, 2000) These concentrate on fitting the student to the institution. Secondly there are those who propose entirely new models. (Berger, 2001-2; Kul & Love, 2000; Rendon et al, 2000; Tierney, 2000) These models include adaptation and institutional change to accommodate the increasingly diverse student profiles.

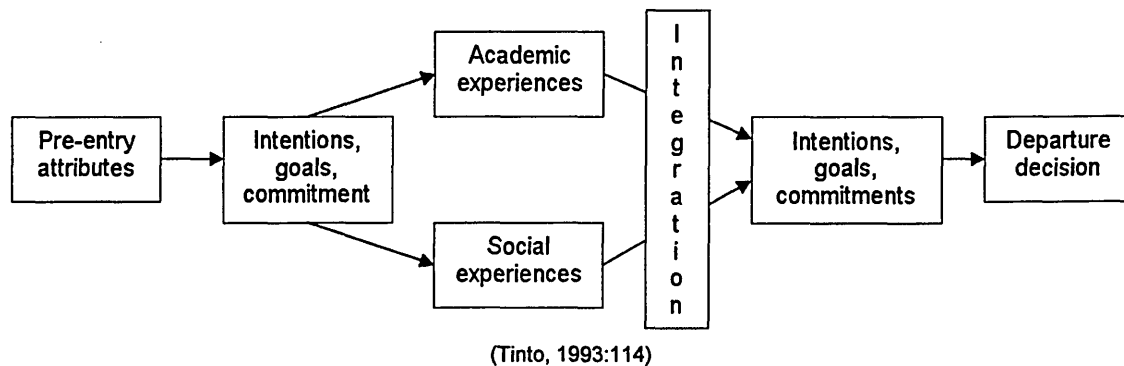
¹ Tinto's model is discussed in paragraph 3.3.2 below

Chapter 3

However, Zepke and Leach (2005:46), who looked at 146 different post-1990 studies, warn us to take care when using a model in a new context. There are often factors that are specific to an institution which may have implication for the transferability of the knowledge and experience gained.

3.3.2 Tinto's Model

There seems to be a general consensus that Tinto's model is the dominant one, so it makes a good place to start. (Moore, 1995) Social and academic integration are key to student retention in Tinto's model. It considers it from a student perspective but "has relatively little to say about the impact of external factors in shaping students' perception, commitment and reactions." (Yorke, 1999:9) Yorke goes on to say that it lacks any emphasis on the institutional contribution to withdrawal.



It claims that student retention can be predicted by their level of academic and social integration. These factors develop over time, as integration and commitment interact. Student Departure (dropout) depends on commitment at the time of the decision.

a) Academic integration

This includes such factors as assessment performance, personal development, academic self-esteem (how well they think they are doing), enjoyment of and studying the subject(s), identification with academic norms and values or identification with their role as a student.

b) Social integration

The level of social integration might depend on the number of friends they have, their personal contact with academic staff or whether they are enjoying being at university.

3.3.3 The Pathway to Retention Model

The Pathway to Retention model was devised by Moxley et al (2001:20-27). It can be adopted by institutions to help facilitate retention. It sets out five objectives to be achieved by the institution of higher or post-secondary education (paraphrased by the author):

1. Perceive a need for retention
2. Establish retention as an institutional aim
3. Expand involvement in retention to support & contribute to the student success
4. Build a retention capacity & establish a formal programme for keeping students
5. Keep students enrolled & working towards their educational aspirations & aims

It goes on to define five forms of Supportive Retention Practices:

1. Emotional support and sustenance
2. Informational support
3. Instrumental support
4. Material support
5. Identity support

3.3.4 Isolating the Main Factors for Attrition

“The evidence shows that there are unacceptable variations in the rate of ‘drop-out’ which appear to be linked more to the culture and working of the institution than the background or nature of the students recruited” (David Blunkett in Thomas, 2002:424)

Factors for Attrition are probably the most documented area of research in this domain. Martinez (1996), though rather dated, offers the best starting point. He looks at a number of case studies in post-16 colleges. The study at Knowsley College, Liverpool, defined a number of student issues that had a real bearing on retention. (Martinez, 1996:16) These are condensed below:

a) Motivation

Factors might be the lack of clear career and/or progression objectives, no particular reason for choosing that institution, resitting the course or transfer after failure from another institution

b) Social

This includes issues such as no friend on same course, lack of support from partner or family or maybe being a different gender or age from rest of group.

c) Time pressures

These might be the result of being a single parent or having young children, caring for sick relatives or being a full-time student with part-time job.

d) Financial

Financial problems might be the result of loss of income support, possible delay of grant/loan, examination fees or daily travel costs.

e) Qualifications

Some students will have only the minimum academic qualification and experience.

f) Any other difficulties

Other difficulties might be unhappiness with some aspect of programme of study, domestic circumstances, health issues, travel difficulties and simply immaturity.

A House of Commons White Paper (2001) concludes that "Information gathered by institutions shows that most students withdraw because of 'personal' reasons or academic failure."

Other factors stated were a lack of preparedness for higher education, changing personal circumstances or interests, financial matters, the impact of undertaking paid work and dissatisfaction with the course or institution.

In an earlier study by Martinez (1995) a number of reasons for attrition (non-completion) were established. These included taking up full-time employment, loss of employment, change of employment, other work related reasons, failed assessment and/or course too hard, obtaining poor grades, dislike of programme, other programme related reasons, family or health issues and financial reasons.

Martinez (1995:6) goes on to say "[a]lthough the precise weight of these factors varies from survey to survey, they are important for all student groups surveyed". Martinez devised the following:

Reasons for coming to college

These included Interest in subject, inability to get a job and preferred college to school.

Source of information about college

Students obtained their information from either college staff or their school.

Support from college

The students felt welcome on arrival in college and felt that they had the support of other students and staff.

Programme issues

Programme issues included lack of satisfaction with amount of tutorial time, being on wrong level of course, not having the correct entry qualifications, difficulties with speed of teaching, programme organisation and programme uninteresting.

Personal and financial

These included receiving help from benefits system, accommodation problems, difficulty paying for bus travel and time off for personal reasons.

Moore (1995) gives a similar list, but added weighting to them:

1. Course unsuitable / disliked	41%
2. Personal reasons	17%
3. Academic problems	11%
4. Financial problems	11%
5. Accommodation	6%

On closer examination of the data Moore discovered that most of the withdrawals occurred during the first part of the first academic year, a large problem being homesickness and feeling lonely and isolated. However, overall most students said they left primarily for course-related reasons. Moore adds a note here and suggests that “students may find it easier to say they left for course related reasons than, for example, acknowledge personal problems.” (Moore, 1995:23) An additional factor found by Moore was that over half of the students surveyed thought that higher education was very different from their expectations.

Yorke & Longden (2004) examine the reasons that students leave their programmes of study. They state that these reasons can be grouped into four main categories. These are flawed decision-making about entering the programme, their experience of the programme and/or institution, failure to cope with the demands of the programme and outside events that impact on the students’ lives.

Post-Script

“[W]e know that it [drop out] is multi-causal, that it is complex and highly context specific, but we also know that it is significantly caused by things which colleges and education centres can do something about.” (Martinez, 1995:23)

3.4 Intervention

3.4.1 Introduction

An institution must give “a commitment to helping students to find their way to higher education, to helping them build the confidence and capacities to persevere towards their educational goals, and to helping them in their education, in their careers and ultimately in their lives.” (Moxley, 2001:1)

When a student withdraws from a course or programme of study there are a number of issues for concern for both the student and the institution. These include loss of educational opportunity, loss of self-worth, loss of higher education places, bad reflection on institution (league tables etc) and financial implications for the institution. (Moore, 1995)

The aim of this section is to review what is currently being done within institutions to reduce withdrawal rates and to provide “appropriate information, guidance and support for students who are considering withdrawing.” (Moore, 1995:4)

As student diversity increases with widening participation, institutions must create a welcoming culture (or ethos) of acceptance, respect and value diversity. (Padilla et al, 1997)

Many more students can be retained if interventions are made at a relatively early stage. This leaves us with the question of what intervention techniques are appropriate and which are the most effective. This is a contentious area and needs careful management. It is undoubtedly multi-causal but is significantly caused by factors that institutions can do something about. (Martinez, 1995)

3.4.2 Improving Teaching & Learning

The whole approach to education from the earliest days has changed substantially over the last couple of decades, and is set to continue to change. Institutions need to be careful to adapt their teaching and learning methods to meet the new challenges. This will be examined in section 3.4.4 below.

3.4.3 Supporting Learning & Supporting Students

A number of studies (Martinez, 1995; Martinez, 1996; Thomas, 2002; Zepke & Leach, 2005) have found that supporting students both academically and

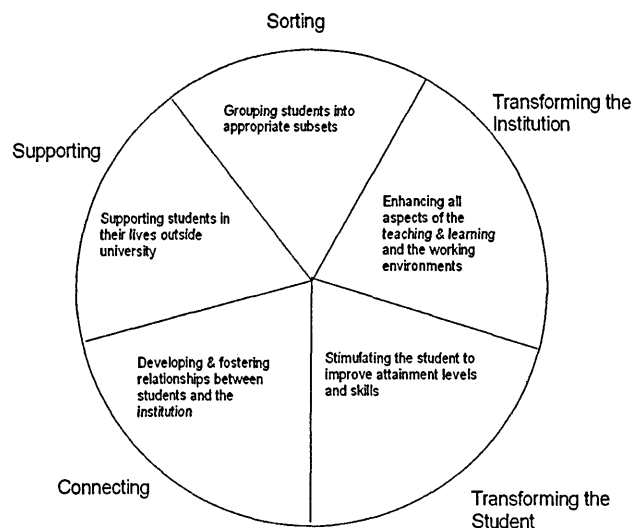
personally can have a positive effect upon retention rates. However, due to the ever increasing student to staff ratio, academic tutors are becoming less able to shoulder this responsibility and are calling on specialist services such as Student Services and the University Counselling Services to fill the void.

Martinez (2005) found that students from high retention courses tended to have more regular tutorials, spent more time individually with tutors and were more likely to rate tutorials as useful. They believed that their tutors had a clear idea of the purpose and value of tutorials and were able to be more precise in what they wanted from tutorials, notably feedback on progress, study skill support, career education and help with progression.

Improved personal counselling and guidance was looked upon favourably but Martinez (1995:22) warns that we are “not yet in the position to demonstrate unequivocally that by increasing guidance and counselling provision we can increase the likelihood that students will complete their programmes of study.” However, there may be some further issues associated with this. Moore (1995:39) found that counselling was generally referred to when students experience personal problems and not as a source to “aid students working through a decision to leave”.

3.4.4 The Beatty-Guenter Retention Strategy Model

There have been a number of Retention Studies conducted in North America. The Beatty-Guenter Retention Strategy Model appears to be a good example. This model divides the task into five distinct areas as shown in the diagram below.



The Beatty-Guenter Retention Strategy Model in (Johnston, 2001:4)

1. Sorting

Sorting strategies include recruitment strategies such as promotion and marketing. They should involve assessment, monitoring attendance and progress and the identification of students 'at risk'. Traditionally universities have used the minimum entry requirements as a method of sorting.

2. Supporting

Supporting strategies are aimed at giving support to students outside the institution's traditional boundaries, to assist them to continue as students. Financial help, child care, health, transport and personal safety all come under this category.

3. Connecting

Connecting strategies are intended to foster relationships between students and the institution. They promote and encourage students into the life of the institution. The use of personal tutors, peer support, faculty advisors and attendance monitoring are examples of such strategies.

4. Transforming the Institution

These strategies aim to improve teaching and learning and the student working environment. This is likely to involve policy changes, curriculum redesign, cultural change, learning and teaching innovations and staff development.

5. Transforming the Student

These strategies attempt to transform the student say from passive learners to active learners or developing and enhancing their skills and expectations. This will involve motivation and the setting of goals. The facilitation of academic and careers counselling are essential here.

Beatty-Guenter (1994) warns that all programmes of retention should include some aspects of all these strategies in order to be truly effective.

3.4.5 Student Diversity

Acknowledging student diversity is a recurring theme. Zepke & Leach (2005) set out their 'Challenges for Practice'. They challenge us to make early contacts (affirmation of students' cultural capital), teach to value diversity and adapt our assessment practices to acknowledge diversity. This research was set in the New Zealand context of Maori and Pacific Island students.

We must not forget student involvement and responsibility in this process. Students still need to address the core academic skills, achieve the requirement and persist to graduation. It is vital that students possess a strong personal responsibility and readiness to retention and any effort to keep students in higher education must respect this in the process. (Moxley et al, 2001)

Student retention has been shown to improve when institutions promote student personal contact outside the formal classroom and show a real commitment to students' total well-being. (Astin, 1993; McInnis et al, 1995)

"Students must come to grips with who they are and where they are going."
(Moxley et al, 2001:5)

3.4.6 Footnote

Moore (1995), surveying 'leaving' students, found that almost 80% of those surveyed withdrew from their courses after having seen staff. Unfortunately we are not supplied with the other part of the evidence. That is, the number of students who sought help and decided to stay. How effective tutor intervention is therefore not measurable.

However, the same survey found that 'Central Guidance' was much more rarely used than academic staff. It also found that the most common issue raised was "links between staff and students ... [they] wanted closer and more regular contact with academic staff" (Moore, 1995:34)

Moore (1995) argues that traditional student support methods based on a personal tutor system have effectively broken down under the pressure of higher student to staff ratios. Tutors interviewed by Moore tended to consider academic issues with students and were far less likely to deal with personal or social matters. Indeed most tutors appeared to have received no formal training on these issues. Comments from both tutors and students showed that both groups were unhappy with the existing situation.

3.5 Success and Barriers to Success

"However, the most important factor affecting institutions' achievement rates is students' entry qualifications." (House of Commons, 2001)

Efforts to widen participation are, to a certain extent being thwarted by some of the larger companies who still insist on a specified number of points at A' level for graduate trainees regardless of the classification of degree the student might have achieved. A' level points and GCSE grades are still being regarded as the

best predictor of future success. Studies consistently show that qualifications at 16+ provide an excellent predictor for continuance in full-time learning. (Johnes, 1990; DES, 1992; Moore, 1995; Kennedy, 1997) "If at first you don't succeed ... you don't succeed." (Kennedy, 1997:15)

As well as the insistence on A' level points and specific GCSE grades some companies insist that all interviewees for graduate schemes should undertake an aptitude test. The author is aware of a student from SHU who worked for a company during his placement year and received excellent reports. He went on to achieve a First Class Honours degree in Business Information Systems. On interview for a graduate scheme at his placement company he was required to take an aptitude test. As a result of his performance on this test he was deemed unsuitable and failed to achieve a place on the scheme.

This practice of taking previous qualifications or additional tests into consideration, or perhaps looking at the institution where the degree was taken is working contra to the widening participation agenda. We must focus on more than "providing opportunities for those who have already achieved to continue to do so." (Kennedy, 1997:15)

The outside perception of universities, 'The Russell Group', 'The New Universities' and the way that the league tables are constructed can have a marked affect on the standing of the degree classification that a student is awarded. It seems apparent that there is a perception by certain employers that a degree from some universities is worth more than at others. Patterson and McPherson (1990) believe that the different ways that universities calculate their figures has more of a bearing on outcomes than the better performance of the institutions.

The subject matter of the specific course that a student takes can have a contributing factor to the student success. Such subjects as medical sciences, education, languages and the humanities have better retention rates than engineering, technology, mathematics and computing subjects. (House of Commons, 2001)

Other issues that were seen as barriers were low aspirations, fear of debt and family pressures, particularly for mature students. (Lall et al, 2003) The debt aversion appears to be a rising problem and may well contribute to the loss of potential students or the early departure of those badly affected. This is a

particular concern for students of working class families who were undecided about entry into higher education in the first place. (Callender, 2003)

“There are many barriers to participation in learning. To widen participation, as many of these barriers as possible need to be removed. Although by no means all of the barriers to participation are put up by providers, they must bear the main responsibility for helping students to tackle them.” (Kennedy, 1997)

3.6 *Student and Institutional Adaptation*

“If a student feels that they do not fit in, that their social and cultural practices are inappropriate and that their tacit knowledge is undervalued they may be more inclined to withdraw early.” (Slack & Casey, 2002:3)

As discussed above, there are two forms of model, those that concentrate on fitting the student to the institution and those who propose models that include adaptation and institutional change to accommodate the increasingly diverse student profiles.

Many of the studies give a great deal of emphasis to the induction of students and to helping them to adjust to the environment of the institution. It has often been taken for granted that this is the major concern. “[S]tudents should adapt to the institution, learning to do things as they are done around here.” (Zepke & Leach, 2005:52) However, “the challenge is to develop ways in which an individual’s identity is affirmed, honoured and incorporated into the organization’s culture.” (Tierney, 2000:219) It may well be that the institutions are set up and function for a particular ‘norm’ of students. The new universities in particular, because they are increasingly catering for ‘widening participation’, may need to examine their structures and approaches to meet differing needs. “A growing number of studies are focusing on the ‘fit’ between the habitus of the student and that of an HEI¹.” (Slack & Casey, 2002:3) Indeed Ready (1998) argues that a lack of fit between student and institution generates anxiety.

Because working class students are less likely to enter higher education there is less experience of higher education in working class communities, and “those that choose to undertake such studies often feel marginalised and out of place in higher education.” (Lall et al, 2003:2) There is also a growing feeling that many students expect universities to be flexible to their needs.

“Students now expect institutions to fit their lives rather than vice versa.” (McInnis et al, 1995, 2000a, 2000b)

¹ Higher Education Institute

3.7 Good Practice

“Clearly good practice needs to mean good business, and changes in the funding arrangements which reward successful collaboration will be required.” (Kennedy, 1997)

Kennedy (1997:91) has set out nine ‘Characteristics of Good Practice’ for institutions to follow. Baroness Helen Kennedy was the chair of a committee that produced the report. The report recognised however, that no institution in his sample exhibited all nine of the characteristics, but several did exhibit some of them. Although these recommendations were devised to enhance widening participation in Further Education, the author feels that they are none-the-less, just as applicable to Higher Education.

Kennedy’s Characteristics of Good Practice

1. Marketing planned and based on intelligence
2. Establish strategies for contacting non-participants
3. Good quality information and guidance, readily available and impartial
4. Have effective support for learning
5. Provide financial and practical support
6. Design a curriculum that is relevant and enables students to progress
7. There is effective teaching and promotion of learning
8. Record meaningful student achievements recognised by employers
9. Have accurate MIS¹ used to evaluate students’ progress

Institutions should encourage students to participate and assist students to stay in learning and to achieve. “These characteristics are more than the sum of their parts.” They need to be seen as a whole and used to underpin and drive organisational culture. (Kennedy, 1995:84)

“Findings should support successful learning.” (Kennedy, 1995:60)

For retention strategies to be effective they need to be implemented at the ‘grass roots’. However, it must be recognised that “[s]ome groups of learners ... are more difficult to recruit, more likely to drop out and less likely to achieve at all stages of their learning.” (Kennedy, 1995:61)

“Education, not retention, should be the goal of institutional retention programmes.” (Tinto cited in Exchange, 2002:13)

¹ Management Information Systems

Chapter 4: What is New in this Research

4.1 Introduction

"Every aspect of college life can be enhanced by trying to improve retention" (Lalgree in Martinez, 1996:37)

Improving retention rates has been the theme of much recent research as outlined in the previous chapter. The author doesn't seek to merely replicate this research but to add a new dimension to the domain. Coming from a Data Analysis background, and having taught 'Business Intelligence' for many years the author feels that his expertise can be used to great advantage in this, attempting to interrogate accumulated data to discover hitherto undetected patterns and relationships in the data.

As the power of computers has become greater and storage space cheaper there has been a dramatic rise in the amount of data that companies store, the vast majority of it in electronic form. "It is estimated that the amount of information in the world doubles every 20 months. What are we supposed to do with this flood of raw data? Clearly little of it will ever be seen by human eyes ... Computers promised fountains of wisdom but delivered floods of data." (KDD, 1995)

Companies have often spent years accumulating this data which they consider to be valuable, but often don't know what to do with it. They know that it probably contains crucial decision making knowledge that could transform their business. As a prominent top manager said, "Who[ever] has information fastest and uses it wins" (Watterson, 1995)

As a response to these trends, the term 'Data Mining' (or 'Knowledge Discovery') has been coined to describe a variety of techniques. "By mining the essential nuggets of information, analysts are able to fully explore existing datasets and identify actionable patterns and trends." (McCue in SPSS, 2003) In brief, decision-making knowledge is held and hidden in the data. Finding and interpreting this knowledge is what Data Mining is all about.

Data Mining tools use a wide range of techniques in order to undertake a variety of tasks, depending on the nature of the problem.

4.2 What is Data Mining

4.2.1 The Concept

Data Mining is “[t]he process of using statistical techniques to discover subtle relationships between data items, and the construction of predictive models based on them.” (Pendse, 2005)



(Kuonen, 2005:3)

That is, Data Mining is concerned with extracting hidden patterns in the data that have hitherto not been found. Data Mining has been developed over the last couple of decades to help explore the vast amounts of data that have been accumulating. The greater power of computers can now be harnessed to enable models (routines) to run at acceptable speeds. The models are based on statistical techniques that are well tested, many of which have been used for many years.

4.2.2 Data Mining as an Inductive Technique

Induction is concerned with the extraction of patterns from the data hence it follows that Data Mining is inductive. (Wikipedia, 2006)

Data Mining might best be illustrated by taking an example in context. A large supermarket chain might use Data Mining in conjunction with their loyalty-card¹ to assist them in marketing their products. Customers can obtain a loyalty card free of charge in return for filling out an application form. On the form there are a number of questions about the customer such as Data of Birth, gender and income. This information can be classified as 'demographic information'. When a customer purchases products at point-of-sale they use their loyalty-card in exchange for discount points. In this way the transaction that has just been created contains two types of information, the demographics of the customer and the products they have just purchased. This information can later be explored using Data Mining techniques.

¹ Store Card that holds Customer Information in exchange for points at point-of-sale

It is the inter-relationship between the demographic data and the transactional data that is crucial to Data Mining. The customers can be classified into similar grouping, such as age groups or income groups and patterns of buying can be discovered.

“There is a story that a large supermarket chain ... did an analysis of customers' buying habits and found a statistically significant correlation between purchases of beer and purchases of nappies (diapers in the US). It was theorized that the reason for this was that fathers were stopping off ... to buy nappies for their babies, and since they could no longer go down to the pub as often, would buy beer as well. As a result of this finding, the supermarket chain is alleged to have the nappies next to the beer, resulting in increased sales of both.” (Fisk, 2006)

This is an example of a Data Mining technique called ‘Market Basket Analysis’ (MBA). It uses the demographic information, gender and age together with the products purchased, in this case nappies and beer to find an association.

4.2.3 Data Analysis Testing

a) Hypothesis Testing

This is the classical statistical approach to quantitative data analysis to assess the statistical significance of the findings. It involves the use of sample data to evaluate a hypothesis about a population parameter.

It involves four steps:

1. State the hypothesis
2. Use the hypothesis to predict characteristics you expect the sample to have
3. Take a random sample from the population
4. Compare the obtained value to the prediction made by hypothesis

This method of testing ‘truths’ in Data Mining is not appropriate as Data Mining is an exploratory means of data analysis (inductive analysis). We do not know what we are looking for until we find it and hence a hypothesis cannot be stated prior to building and running the models. Another form of testing is needed in this sort of situation. The one specifically developed for this purpose is Exploratory Data Analysis (EDA).

b) Exploratory Data Analysis

As opposed to traditional hypothesis testing EDA is used to identify relationships between variables when there is little or no knowledge of the nature of those relations. Typically in exploratory data analysis many variables

are taken into account and compared, using a variety of techniques in the search for systematic patterns. The process is gradually refined by removing variables that are of lesser significance. This process is known as variable reduction. It involves the three stages of exploration, model building/validation and deployment.

Stage 1: Exploration

This involves the process of data preparation; that is the cleaning and transforming the data. Subsets of the data are selected and some pre-processing will be done in order to remove obvious redundant variables hence reducing the variables to a more manageable number. Data mining sets can have a very large number of variables. A wide range of graphical and statistical methods may be used.

Stage 2: Model Building and validation

This involves examining various Data Mining modelling tools and choosing the most appropriate ones. Each one will have a measurement of predictive performance such as Support which is a measure of the percentage of data-points that possess the predicted pattern and Confidence which is a measure of how certain the pattern is believed to be true. A support of 25% and a confidence of 80% would mean that 25% of the data displays the given pattern and the model believes, with the data evidence that it has been given, that it is 80% certain that it is true. It is usual to use many different modelling tools with a variety of different settings before settling for the optimum ones.

Stage 3: Deployment

Once the models have been created and tested the final stage involves using the models as built in Stage 2 and applying them to new data in order to generate predictions or estimates of expected outcomes from this new data. (Statistica, 2006)

These stages will be looked at in more detail in Chapter 8, 'A Methodology to Use'.

4.2.4 Underlying Activities

Data Mining can be used in a number of different ways. Berry & Linoff (2004) define these as classification, estimation, prediction, affinity groups, clustering and description.

a) Classification

This is the process of examining a newly presented object and assigning it to a specific category in a pre-defined list. For instance someone of age fifteen might be classified as in the 'teenager' category.

b) Estimation

Estimation is an approximate calculation of a result, often based on previous knowledge and other factors. For instance you might estimate the value of an item by looking at such things as age, size and substance it is made of. Experience of the value of similar items would assist greatly in this task.

c) Prediction

A prediction is judging what will happen in a given circumstance or situation. For instance you might predict whether a customer will CHURN¹ in the next three months.

d) Affinity Groups

An affinity group is a collection of objects that share similar traits or values. They are in some way associated with each other. For instance, items that are bought together at a supermarket could be considered to be an affinity group. (MBA²)

e) Clustering

This is the process of forming like groups with similar characteristics. In a university context it might be to do with similar demographic features of students and/or the subjects they are studying.

f) Description

Description or visualisation is often referred to as exploratory or visual data mining. It is designed to provide strategic insights from the data and guide future decision making.

¹ CHURN is a term derived from 'change' and 'turn' and means a customer moving to an alternative supplier

² Market Basket Analysis

4.2.5 Types of Learning

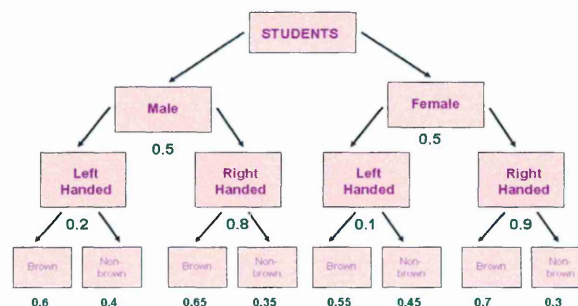
a) Supervised Learning

This is where the model is given an outcome to predict. For instance, predicting the amount of money a customer might spend in a supermarket. You have the demographics of the customer and you have the previous spending patterns for that customer and other similar customers. From these, using an appropriate model, Data Mining can predict, with a reasonable certainty, the amount of money the customer might spend.

Classification, prediction and estimation are supervised activities, since they are all looking for a specified outcome. Decision Trees and Neural Networks are examples of models that can be used for this purpose.

Decision Trees

A decision tree is a table of decisions and their possible consequences used to create a plan and reach a specific outcome. Decision trees are used to assist decisions making. For instance, a group of students are surveyed and divided up by gender, handedness and eye colour. The diagram below shows the possible outcomes. From the diagram below we can deduce that there are equal numbers of male and female students in the group. However, the probability of left-handedness amongst males is 0.2 whilst that amongst females is 0.1¹. There is also a differential in probabilities for eye-colour amongst the different groups.



A Decision Tree

Decision Trees then, are techniques that can be used to predict future outcomes and identify factors/variables that can be used as predictors for these future outcomes. They give interpretable rules and logic statements to help in the decision making process. A decision tree processes a series of input variables and partitions the data into smaller and smaller segments termed

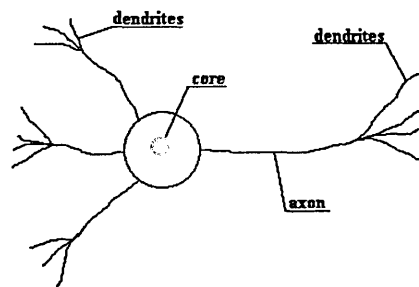
¹ This data is fictitious

nodes in its pursuit of the target or output variable. This partitioning continues until it reaches the pre-defined stopping criteria. Any leg of the partition that has a probability of less than a preset value (say 0.01) at any stage will be 'pruned', that is, stopped. A weakness of this model is that the variables are taken in a specific order. The order that the variables are taken could, in certain circumstances miss patterns of behaviour.

Both SPSS® and SAS®, the principal Data Mining software vendors have produced a decision tree algorithm called C5.0.

Neural Networks

A neural net seeks to artificially replicate the human brain in its learning process. The term 'artificial' means that neural nets are computer programs that can handle a large number of calculations in its learning process. The human brain has more than a billion neural cells that process information. Each cell is a simple processor; the interaction of these in parallel processing makes the brain able to function so efficiently. (Etimage, 1997:2)



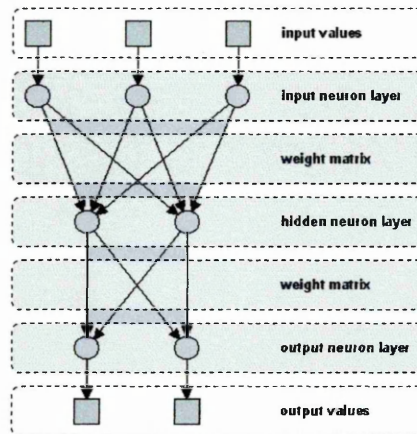
Structure of a Neural Cell in the Human Brain (Etimage, 1997:3)

A neural network takes a series of inputs and a known output and is trained with data where the value of the known output is available. It can then be used in a new situation to predict the output. Inputs might be such things as age, gender, ethnic origin, social class and education with a possible output of salary.

Once the inputs and output(s) have been determined the network needs to be trained. This is done by obtaining a large amount of data where the output(s) is available. Once a pattern of inputs has been established then predictions can be made using data where the output is unknown. The diagram below shows how artificial neurons can be connected together.

Neural Networks are probably the most used of all Data Mining techniques as they are straightforward to use and relatively stable with large datasets. They can be used in similar circumstances to Decision Trees. However, Decision

Trees work best with a small number of input variables but the built module becomes very complicated as the number of variables increase. It has the advantage over neural networks as it has interpretable rule based outcomes which a neural network does not have. However, "Neural networks work well with datasets containing large amounts of noisy [unclean] input data". (Roiger & Geatz, 2003:256)



Neural Net with three neuron layers (Etimage, 1997:5)

Neural Networks can be used by novices and experts alike as they have a set of default parameters that can be checked and amended by the more experienced user. There is a danger of over-fitting the data so the models performance is constantly assessed against validation data. Over-fitting is the process of using too many variables (co-variants) to predict the outcome. It can lead to spurious or incorrect associations.

The diagram above shows how the input variables (values) are interconnected to establish associations in the data. Each node in the neural network is referred to as a neuron to mirror the workings of a human brain. Each neuron is like an on-off switch. It either accepts the connection or rejects it according to a threshold value.

The input variables are fed into the neural network without any discrimination. This forms the input neuron layer. However, not all the input variables are of equal value hence they are weighted in the next layer proportional to their relative importance. The input from the weight matrix is processed by a propagation function that combines the values of all incoming weights. This is then passed on through one or more hidden neuron layer with their corresponding weighting matrix until an output or outputs are determined.

An advantage of a neural network over a decision tree is speed when working with large data sets, particularly with noisy (unclean) data. A disadvantage is that the process is hidden from the user. That is, the user cannot see what has happened inside the 'black box' of the neural network as no rules are given. (Roiger & Geatz, 2003)

There are neural network algorithms in both SPSS Clementine® and SAS Enterprise Miner®.

"While it is easy to explain decision trees NNs [Neural Networks] are much more difficult to understand". (Dunham, 2003:105)

b) Unsupervised Learning

Unsupervised learning is where the model itself finds the outcome. Instead of asking the model to predict a specific outcome, the model is asked to find a relationship or pattern in the data. Affinity grouping is an unsupervised activity as is clustering. The study will mainly concentrate on these. They will be explored in Chapters 8 and 9.

Rule Association

Affinity grouping, or rule association is the basis of Market Basket Analysis. A series of inputs are taken and inter-related patterns are sought. For instance:

If you are in a pub and you buy a pint of beer and don't buy a bar meal, you are more likely to buy crisps at the same time than somebody who buys a bar meal.

Rule Association looks for interconnections in the data and attempts to examine cause and effect whilst establishing a degree of certainty. There are usually two test parameters. The first one is 'support'; that is what proportion of the data support the rule. Here we might be looking for a percentage of around 20-25%. The second parameter is 'confidence'; that is how confident the model is that the conclusion is true for the population that it is drawn from. Here we might be looking at a certainty of say 70-80%. If you start off with a high support and high confidence you will get few if any rules. The more that you reduce these parameters the more rules you will get, but the less confident you will be that they hold true for the population in general that you have sampled by your dataset. These rules are used to make predictions or estimates of unknown values or variables.

The rules are given in the format of:

if <condition 1> and <condition 2> then <conclusion>.

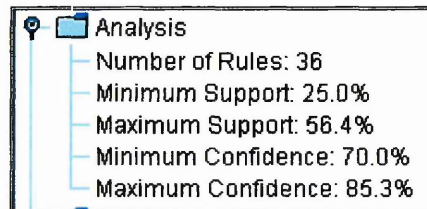
An example might be:

IF student stressed AND student in debt
THEN student leaves

It might be expressed in the form of antecedents and consequent(s):

Antecedent 1	Antecedent 2	Consequent
Student stressed	student in debt	Student leaves

In the table below an Apriori model (rule association) was built with a preset minimum support of 25% and minimum confidence of 70%. Once the model had been built to the preset parameters it established thirty-six rules and gave a maximum support of 56.4% and maximum confidence of 85.3%. It should be concluded that this was a stable and reliable model as the rules were created within our pre-stated parameters and established support and confidence far higher for at least some of the rules created. Indeed had a model been built with a support of greater than 56.4% (say 57%) and a confidence of greater than 85.3% (say 86%) then no rules would have been created.



Minimum support and confidence should be declared before any model is built. A series of models are then built that have varying levels of support and confidence on or above the pre-declared levels. The optimum model will be the one that gives a manageable number of rules that are on or above the pre-set level. The thirty-six rules it found above is probably too many to manage with a danger of the model over-fitting. (See 4.2.5 above)

SPSS Clementine® and SAS Enterprise Miner® have produced two different algorithms to be used for rule association. They are GRI (Generalised Rule Induction) and APRIORI (A Priori).

Clustering



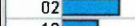






Clustering is dividing data up into groups of similar characteristics. This is useful when looking for trends. It gives the end user a high level view of the data. However, any reduction in detail loses richness in the data and hence some of the patterns that might be established. It does, though, simplify the data and give you more manageable data on which to work. Historically clustering techniques have been used by mathematicians and statisticians for a

significant time period. An example of clustering within context might be to consider the demographics of students. Clustering in this context would refer to the process of organising them into groups according to similar characteristics, such as age group, gender, religion and ethnic group.

In an ideal case all the data-points within a cluster would have identical values for the predictors they were clustered on. However, in practice this is rarely the case, particularly when there are a large number of predictors. There would turn out to be far too many clusters, some of which would have few data-points in them. It is then, a matter of compromise. The number of clusters produced need to be of a reasonable number and contain an acceptable number of data-points in each.





Many of the clustering algorithms such as Kohonen Networks (an unsupervised Neural Network) and K-Mean allow the user to choose the number of cluster to look for. However, in practice it means running the model a number of times with differing numbers of clusters until a stable group of cluster are produced. Even when this has been done, some of the cluster might be ignored since they contain too few data-points.

The example below is output from a Kohonen Network which takes a matrix approach to cluster creation. Each group of clusters are formed on a 2-dimensional grid. The example below was created on a 3x3 coordinate grid with 0 to 2 on each axis:

Val...	Proportion	%	Count
00		19.93	117
01		7.5	44
02		14.48	85
10		9.03	53
11		2.73	16
12		6.64	39
20		15.16	89
21		7.67	45
22		16.87	99

It can be seen that four of the clusters are significantly more populated than the others. In this case there are reasonable grounds to reduce the model to only those clusters. This does, of course reduce the number of data-points being considered, but those that are used are relatively stable.

The distribution below shows the reduced number of clusters.

Val...	Proportion	%	Count
00		30.0	117
02		21.79	85
20		22.82	89
22		25.38	99

The original dataset had 587 data-points, when the number of clusters was reduced the number data-points used was 390, that is 66.4% of the actual number. In large datasets this should not be a problem, but the analyst needs to be aware that the constructed model is a compromise. The number of clusters to concentrate on is a matter of combining experience (tacit or implicit knowledge) with solid hard investigation (explicit knowledge).

There are two well-known clustering techniques that have been programmed in Data Mining. These are Kohonen Networks and K-Mean (MacQueen, 1967; Hartigan & Wong, 1979). Again SPSS Clementine® and SAS Enterprise Miner® have versions of these algorithms.

4.3 Data Mining in Industry

“Forward-thinking companies today are using data mining to reduce fraud, anticipate resource demand, increase acquisition and curb customer attrition.” (SAS, 2006b)

There are two principal Data Mining software vendors, both of which are also the major suppliers of statistical software. These are SPSS, with their Clementine® software and SAS with their Enterprise Miner® product. They have similar functionality.

SAS (2006a) summarise the main Data Mining tasks as:

“Seek and retain your most profitable customers”

Use customer demographic information together with their buying patterns to develop lasting relationships with them by anticipating and fulfilling their future needs.

“Segment markets for a targeted approach”

Implement target marketing to increase response rates by only targeting the customers that are thought to be the most likely to wish to purchase the commodities.

“Predict the future and identify factors to secure a desired effect”

Improve the quality of the production process by anticipating problems before they happen, forecast stock demands and assess the risk of applications for customer credit.

As mentioned above Data Mining is used extensively in the retail trade to help predict future demand and customer buying patterns. It can help in store layout, suggesting which products should be placed next to each other. It is also used extensively in the pharmaceutical industry.

4.4 Data Mining in Higher Education

“Data mining is a powerful tool for academic intervention. Through data mining, a university could, for example, predict with 85 percent accuracy which students will or will not graduate. The university could use this information to concentrate academic assistance on those students most at risk.” (Luan, 2004)

There is very little written about Data Mining in Higher Education, and much of what has been found appears to have been written by or in conjunction with the software vendors themselves. However, these sources should not be ruled out as being biased since it is often through partnerships with software vendors that innovative data mining has taken place.

SPSS are a major software provider for Data Mining. Their Clementine[®] suite is the author's chosen software for this study. An SPSS Executive Report written by Dr Jing Luan (Chief Planning and Research Officer, Cabrillo College¹) gives a great deal of food for thought. He looks at the possible use of Data Mining in Higher Education and tries to compare it with comparable uses in Industry.

Higher education faces many challenges, such as predicting the paths of students to graduation. Many institutions would like to know which students will need assistance in order to graduate and the sort of assistance required. Some students may be more likely to transfer than others. Data mining may be able to warn an institution to take action before a student drops out, or to predict the number of students that will choose to take a particular course and allow resources to be allocated efficiently and effectively. (Luan, 2004:2)

As discussed above, the author believes that unsupervised modelling is likely to be most beneficial to this study.

“Unsupervised data mining is often used first to study patterns and search for previously hidden patterns, in order to understand, classify, typify, and code the objects of study before applying theories.” (Luan, 2004:3)

A graduation database holds many records about students which, when student demographic overlays are placed on it would yield a wealth of information. Data Mining would allow a model to be built from the graduation database, then “the analyst can feed in another student group, such as new students; the model applies the learned information to the new group to predict the likelihood of graduation.” (Luan, 2004:3)

¹ Cabrillo College is in Aptos, California

Below are a number of Private Sector business questions and their equivalent Higher Education questions.

Private Sector Questions

Who are my most profitable customers?
Who are my repeat Web site visitors?
Who are my loyal customers?
Who is likely to increase his/her purchases?
Which customers are likely to defect to competitors?

Higher Education Equivalents

Which students are taking the most credit hours?
Which students are most likely to return for more classes?
Who are the "persisters" at my university/college?
Which alumni are likely to make larger donations?
What type of courses will attract more students?

(Luan, 2004:4)

Prof Michael Hardin of the University of Alabama, with the help of SAS® has developed a student retention model that identifies new students who are at-risk and provides results early enough for the university to intervene. The model sifts through enrolment records and standard 'freshman' surveys to identify key variables that affect retention. He states that the University of Alabama's retention rates have risen as a consequence. "As the retention rate increases, the university's rankings go up, and everybody's degree becomes more valuable." (SAS, 2006c) However, it appears to use only demographic information such as parental educational levels and high school grades without mention of student personal problems.

4.5 Summary

In this chapter the subject of Data Mining has been discussed and put into the context of the subject area under investigation, namely student retention. There are many data mining techniques that could have been investigated. However, only those appropriate to the study were outlined. This is not to say that these are the only ones that could have been used. However, as in most studies, choices have to be made for expediency and the most appropriate ones chosen. Considerations were based on the suitability of the technique, its ease of usability and the author's familiarity with it, given that there was little choice between the options.

Chapter 5: Rationale of Approach

5.1 Rationale

The purpose of this study is to critically explore the issues that affect student non-completion and what measures may be taken to increase retention rates within the Higher Education Environment. It examines current literature and studies on this subject matter focusing on those in similar environments to SHU.

5.2 Research Methods

It was decided therefore that the research would take a mixed methods approach, using the most appropriate techniques at each stage. The use of mixed methods implies the combining and integrating of qualitative and quantitative research methods. (Burch, 2003)

“Using a ‘mixed methods’ approach, (a combination of quantitative and qualitative techniques) can produce a much deeper understanding of your research question. It is also looked upon favourably by potential funders.” (CEM, n.d.)

The research falls into three stages, the first is student centred and seeks to locate and understand the issues affecting retention (Objectives 1 to 4). The second part is associated with tutor intervention (Objective 5). Finally recommendations will be compiled to suggest how data mining techniques can be used to greater advantage in the future (Objective 6).

In order to conduct the research, a number of assumptions need to be made, particularly about the students that are the subjects of the research. HEFCE¹ defines non-completers as those students “who are deemed not to complete their studies in the period for which they have registered.” (In McGivney, 1996:23) This implies that they are deemed to have completed if they get as far as the end of year examination and sit it, regardless of their result. Those dropping out and undertaking another undergraduate programme of study will not be deemed to have failed. An article in the Times Higher Education Supplement (THES, 1995a) suggests that an increasing number of these students aren’t dropping out; they are shifting around.

The research firstly assumes that the students are able to understand and respond to certain lines of discourse. Since HE students have to achieve a relatively high standard of entry qualifications and achieve a recognised standard in English Language it is not an unreasonable assumption to make.

¹ Higher Education Funding Council for England

Secondly the research assumes that a student's failure does relate meaningfully to things that can be described. That is, the issues are understandable and tangible. A lot of research has already been done in this field over the last few years. The author will seek to draw upon the results of this research and extend it in order to make sense of the findings.

It is further assumed that there is sufficient commonality between students and their issues to be able to abstract and generalise in order for recommendations to be made. The commonality of issues will be discovered by using Data Mining techniques.

Lastly, it is assumed that there are a number of retention issues that are well documented in other research. Other issues may arise because of new circumstances, such as increased student debt. These issues will change over time and might indeed grow in importance or diminish as solutions are found.

5.3 Student Issues

From background reading it has become apparent that there are a number of issues relating to student retention that are well documented and are in the public domain. The author has derived these issues from literature and seeks to explore and extend these with a qualitative study within the ITPA in SHU. What follows is a series of in-depth semi-structured interviews with a small number of students who either failed or were likely to fail at least one module in this academic year (2004-5) or the previous academic year (2003-4). This initial research was aimed at verifying the major issues that contribute to failure and non-completion that have arisen from the literature and identifying any additional issues that have not as yet been identified previously. This approach may be deemed to be thematic. Semi-structured interviews are useful as exploratory studies and a prelude to quantitative study. They can provide in-depth material to enhance a large-scale survey. They use "a schedule of questions very much like a questionnaire. The questions are usually open and the responses should be taped for later transcription". (University of Hertfordshire, 2000) This method is used in qualitative studies for the analysis of the focus group interviews or open-ended questions. These first two research exercises (literature review and student interviews) are aimed at finding answers to Objective 1: *"Identify from literature and secondary research, the student issues that contribute to dropping out"*.

Once these issues are established the research seeks to establish how general the problems are within the wider research population. The research analyses to date are formulated into a series of questions to be asked in a survey. The administration of the survey is by the use of a questionnaire. "The questionnaire is a widely used and useful instrument for collecting survey information, providing structured, often numerical data." (Cohen et al, 2000:246) It is formulated and administered to a much larger sample of students, drawn from the research population of SHU and a number of similar universities. No restriction is placed on this research to include only those that have been deemed to be failing, but it is restricted to computing students. A member of the computing academic staff with an interest in student retention was identified in each of the other universities to act as contact. The aims and rationale of the study was made clear and the questionnaire explained. They became the vehicle to ensure a good response of students. The questionnaire was available on-line. This method of administering the questionnaire had two clear advantages over traditional methods of delivery. Firstly that of greater student coverage and secondly the results could be tabulated automatically thus saving a huge amount of time. Response rates to a questionnaire have long been a problem. "Respondents cannot be coerced into completing a questionnaire. They might be strongly encouraged." (Cohen et al, 2000)

5.4 Tutor Intervention

Once the results of the interview analysis became available had been the issues identified, tutors were then consulted. A focus group of staff was convened, drawn from academic teaching and pastoral staff. Their brief was to find their views on what tutor intervention can help alleviate the effect of the issues identified as contributing to student failure. A focus group is a qualitative research method which involves a group interview conducted by a moderator and uses a discussion guide written to meet the needs. It usually requires the participation of around eight. (Airms, 2004)

Much attention was placed here on the interrelationship between the issues. This process was used to provide enlightenment to Objective 5, *"Identify by the use of an ITPA tutor focus group, where tutor intervention can be used to help alleviate these issues"*. Again the data was analysed by the use of 'Content Analysis' techniques. (See above)

5.5 Recommendations

The recommendations arise out of experience gained from applying data mining techniques to higher education retention issues and how these techniques might be used to greater advantage in the future. This focuses on Objective 6, *“In the light of the findings compile a list of recommendations on how the use of data mining techniques could be further developed”*.

In addition a series of recommendations aimed at improving student retention will be drawn up in conjunction with the focus group mentioned in Section 5.3 above and put back to the group for consideration and approval. Jefferson College (2002) in Kentucky USA produced a list of sixty-three ideas subdivided into four categories. These ideas seem to have attracted a lot of interest. The categories are Faculty/Student Interaction (16 ideas), Classroom Management (22 ideas), Student-Initiated Activities (6 ideas) and Faculty-Initiated Activities (18 ideas). The ideas range from such fundamental issues as learning students names to the encouragement of students to set personal goals.

5.6 Ethical Issues

In a survey of this sort a number of ethical issues are likely to be encountered. Since students are asked about things that are currently affecting them, then the issue of prying into a student's personal life might well be raised. This issue needs to be handled sensitively. All interviews were tape recorded to allow the interviewer more freedom to explore the issues without the need for notes being taken. However this increases the need to guarantee confidentiality. Cohen et al (2000:62) says that “although researchers know who has provided the information ... they will in no way make the connection known publicly”. All tapes were destroyed after the interviews have been transcribed and all responses will be anonymised. This procedure was explained to those being interviewed.

Students and staff were assured that their views expressed will be treated in absolute confidence and that it will not be possible to identify them from the context of the research. Each person agreeing to be interviewed was asked to sign an agreement outlining the use that the finding would be put to and guaranteeing their anonymity. It also asked for their agreement to use their expressed views by means of a quotation. In addition, they were told that they could withdraw from the research at any time should they desire to do so up to

the point of publication. Since the students answering the questionnaire were anonymous then the issue of confidentiality was not applicable in this context. The questionnaire was online and hence there was no possibility of identifying students from the point of initial entry of responses.

In addition SHU Computing attracts only about 10-15% female students. In a national survey, Rogerson (1997) found that only 12% of Computer Science graduates were women. Although this may well have improved a little in recent years it appears to be stubbornly low. However, the analysis of “females in computing” is not part of this study; neither is the disproportionately large number of students of Asian descent. Both these issues could be explored using data mining techniques.

5.7 Educational Issues

In the latter stages of the study, when examining tutor response, it was necessary to assess how tutors can intervene whilst still maintaining educational standards. This was tackled by the formation of a ‘focus group’. The nature of this group was crucial. A cross section of tutors were needed to be found that were willing to participate. The size of the group was no larger than eight so as not to limit discussion. (Airms, 2004) Care needed to be taken in the selection of the members of the focus group, as Morgan (1997:35) points out, “a randomly sampled group is unlikely to hold a shared perspective on the research topic and may not even be able to generate meaningful discussion.”

However comprehensive the recommendations might be and however desirable they might be deemed to be by the management and tutors alike, those tutors who claim that “it’s not their responsibility” will inevitably limit the success of the exercise.

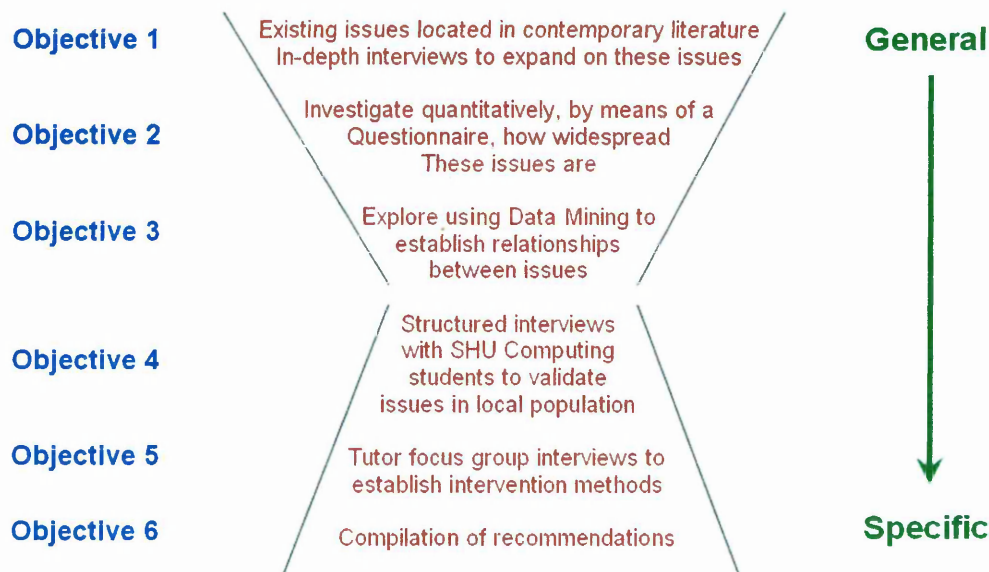
5.8 General Problems

The research started by examining students who were identified as “failing” in the January student progress review. Unfortunately this missed the students who had already dropped out. These students were not contactable for questioning. There is no real way round this problem as the January review is the first time that statistics have been compiled for discussion. Students who have not dropped out but attend infrequently and hence more likely to drop out are extremely difficult to contact.

The study is sensitive to the issue of ensuring honesty of response. One to one interviews and guarantees of anonymity help to reduce this. However, it is noted that some students might want to supply the answers that they think the interviewer is wanting rather than their own genuine views. Every effort was made to limit the likelihood of this form of response. "It is important for the interviewer to be able to explain how the particular respondent came to be selected for the sample and why it is important that he or she, rather than someone else, takes part." (Oppenheim, 1992:82)

5.9 Research Overview

Below is set out diagrammatically how the different research elements fit together along with the objective that it is attempting to cover. This shows that the research starts with the general (Objectives 1 to 3). It started with the issues that have been identified from a comprehensive literature review and backed up and extended by in-depth semi-structured interviews. Once the generality was established the research moved into the specifics of SHU ITPA, (Objectives 4 and 5). As a result of the study and the examination of data mining techniques in the student retention context, then a series of recommendation will be drawn up to show how these techniques can be used to greater advantage in the future (Objective 6).



Chapter 6: Initial Interviews

6.1 Introduction

After reviewing the available literature on Student Retention a number of categories were drawn up. These categories rely heavily on the work of Woodley et al (1987), Tinto (1993), Long (1995), Yorke (1999) and Yorke & Longden, 2004); see section 7.4 below. The final interview list was defined as:

- Course
- Institution
- Academic
- Personal
- Motivational
- Financial

An interview schedule was devised based on these categories. This can be found in Appendix B. Six computing students from Sheffield Hallam University who were identified as failing to achieve were randomly selected from various years of study and interviewed. This was done to confirm the major issues and to gain clarification and insight into these issues. Sample interview transcripts can be found in Appendix A.

6.2 Course

The first category to be explored was the course itself. Most of the Students interviewed had chosen the course well in advance and were confident that it was what they wanted to do. However some had doubts about whether they had the right skills to undertake it. "Is it too difficult? Is it above me kind of? Can I do this? ... Will the course lead me into a job that I want?" (Student A)

A couple of students had started late and found significant problems at the beginning. One transferred from another course within the same university and another came from elsewhere. This led to a significant mismatch between expectation and actuality, "I found it different to what I read." (Student C) In particular Mathematics was cited. "I didn't realise that there was Maths at first and how hard that would be." (Student C) The student went on to say that a couple of their fellow students had dropped out because of the Maths; this was echoed by other students. Programming was also found to be a particular problem. One had to seek out a tutor from another university for help during the summer vacation after a referral in the May examinations. "I needed a bit of

one-to-one sort of tutoring, instead of sharing the issues with the whole class.” (Student C) The student didn’t feel able to get this from SHU tutors. This was reiterated by Student E who said that when tutors were eventually tracked down the amount of time given was inadequate.

Some students expressed doubts about the relevance of some of what they had been taught, particularly in the early years. One student went as far as to say “from the way we did it, it just didn’t seem like it would ever get done [like that] in an organisation”. (Student F)

Adapting to the university ethos was a common theme with a number of those interviewed. “I found the first year difficult to adapt [to] coming straight from college.” (Student D) Issues such as the amount of time to spend on a piece of work were cited as examples. One talked of a culture shock after still being treated like a child in sixth form and expected to be a mature adult in university. (Student F) The differential backgrounds and aptitudes of Students also caused concern. One expressed consternation at “doing what you’ve done in God knows how many weeks and they do it in a couple of hours!” (Student C) Student D stated that in the first year, if you did the appropriate amount of time on a piece of work you would get high marks regardless of content.

6.3 Institution

The second category examined was that of the institution itself. It was felt that the location of the university was of primary importance. “Location, I think is fantastic.” (Student C) The university being in the centre of the city was felt to be a distinct advantage, though student accommodation was often a fair distance away. “The university is right in the centre, easily accessible from anywhere.” (Student E) Home students did feel disadvantaged socially but didn’t feel disadvantaged academically.

The primary issue was that of facilities, their quality and their availability. All had had problems at some stage citing available work space in the library, particularly in the quiet area as being of particular concern. “I had to get up really early and get there before 9 o’clock to get a computer in the silent area.” (Student C) The problems were felt to be particularly acute towards the ends of term where things like logging into computers could take up to quarter of an hour.

It was felt that in general, the student accommodation was quite good and that it was necessary for first year students to be housed in halls of residence. "You're not quite as independent as you are in the second or third year." (Student F) It was thought that everyone was in the same situation when they first came to university and that having meals provided and paid for in halls was a distinct advantage. However, all the non-home students appreciated living with friends in student housing in later years.

6.4 Academic

By far the most recurring academic theme was study skills such as organisation, essay writing and note taking. It was thought that these could be tackled "like we had professionalism and communications class in our first year. We never covered note taking or anything like that". (Student C) Essay writing appeared to be a particular problem, chiefly considering the type of A' levels that many of these students had studied.

Formal examinations were also considered to be a significant problem. "Examinations, I'm a nervous wreck." (Student C) Such things as not being able to sleep the night before examinations and a general fear of examinations were common. Student E stated that panicking about getting to the examination hall in time was a concern, particularly if it was in another part of the city.

The student perception of the first year not being important was discussed. "Everyone says when you come to university you don't have to go to lectures in your first year, it doesn't matter, it doesn't count for anything, as long as you pass you're alright." (Student F) This sentiment was generally agreed to be the norm initially and led to bad attendance at lectures and tutorials. It was accepted to be a major contributing factor to student drop-out. This perception was modified radically as they progressed through their courses.

Most had tried not to get behind with their work, but felt that if they did it was very difficult to catch up. It had often resulted in a module being referred or even failed outright. However, it wasn't considered to be the only reason why a module was failed. Bad or inadequate teaching or tutor support that was below expectation was cited as significant, "some tutorials they didn't know what they were doing! I don't know where they had been pulled from". (Student C) Late

return of submitted work was felt to be a problem. Student D stated that they were sometimes treated as if they didn't have a right to know about things. Most agreed that they had had undue criticism from tutors and/or peers at some stage.

6.5 Personal

A range of personal issues were outlined in the interviews but by far the most significant was balancing the need to study against the need to earn to finance these studies. All the students interviewed had had part-time jobs at some stage in their university life. It was felt that parental support and a student loan was insufficient to meet their needs. For various reasons parental financial support wasn't always forthcoming. "I funded myself through university." (Student F) The issue of finance is discussed later.

Problems with close relationships (girl friends and boy friends) rated highly with most of the students, particularly when they were living in the same house. Student A said "my coursework was rushed and my exam revision was more rushed". The student rated this as the greatest problem. The break up of these relationships was cited as having a significant effect on their ability to concentrate and consequently their work. It was highlighted as being a contributory factor in failure to complete work and subsequent module referrals and failures. Closely connected with this was the issue of socialising. It was apparent that some weren't willing to sacrifice friends and social life just to get a high degree classification (Student B), particularly in the first and second year.

Student C had started a course at another university but had transferred to Sheffield because of not being able to settle into the life of that university. The majority of the students that this student had associated with had been home students with friends outside university. However this had not been a concern after transfer to SHU.

Parental pressure appeared to rank highly. Most students had encountered significant problems at some stage ranging from distractions when working at home (Student A) to having to look after aging parents that didn't speak fluent English. (Student D) Student C's father had been a fitter down a mine which it was felt had led to "old fashioned" parental values.

There were a number of deeply emotional issues that some of the students had encountered. Student C spoke at length about the problem of admitting being gay to his parents and friends and how this had caused a great deal of stress to the student for a long time. It was felt that it may have affected his studies. He cited this as the main reason why he had moved out of his parents' home into student accommodation much against the wishes of his parents.

Alcohol had not been a serious problem, but had caused some of the students to miss the occasional lecture or tutorial. (Student F) Most students said that they had friends that had drug related problems but it hadn't really affected them. However, Student E did admit to going to raves for a time where drugs were being taken. This was described as being sort of scary.

A number of the students admitted to having self-confidence problems and even depression. "I think that [lack of self-confidence] really did have a bad effect on my grades ... I just wasn't happy with myself and I wanted to sleep all the time." (Student F)

One student had an accident part way through the first year which resulted in a significant amount of time in hospital and at home recuperating. On return to university it had caused two significant problems, that of catching up with the work which was serious, and the loss of friends. "[W]e just drifted apart ... I was really worried and when I came back and saw the workload I didn't think that I was going to be able to do it." (Student E) Help was sought from a tutor who gave assistance on how to park issues and deal with them at a slower speed.

6.6 Motivational

Motivational issues were wide and varied. Student A spoke of not doing very well at A' levels and as a consequence had a fear of failure. This was described as the student's major driving force causing continuous tension and stress. However, the student did admit to being switched off if the work became too hard. A number admitted to having had attendance problems, particularly in the first and second years which was a major contributing factor to module referral or failure. "My attendance has not really been always very good" (Student B), the one who was not "willing to sacrifice friends and social life just to get a first".

Peer group pressure appeared to be a strong demotivating influence. "[A] lot of my friends pester me to go out when I'm working." (Student C) However, they

agreed that the influence was less as they approached their final year. "I've had the student life at the beginning and now I'm not wasting four years. I don't want to come home without a degree, a good degree." (Student E) This sentiment was reiterated by others. Student D was the youngest child (ten years younger) of a large family. All the older siblings had graduated from university so competition was expected. "Just generally I've wanted to do well at whatever it is." (Student D) At times peer group pressure appears to be a motivating influence, particularly competition to do well in assessment. Group work, however, has been cited as a strong demotivating influence, "because if you don't get into a good group it is going to affect your grades." (Student F)

The strongest motivational issue appears to be the industrial placement year. It was generally considered to be a "maturing thing". (Student F) Those students who had undertaken placement in their third year had come back to university determined to do well. "It's a matter of getting my head down ... I'm very much looking forward to going back." (Student E) Coupled with this was the thought that achieving would help the students to find a "job associated with money". (Student D)

6.7 Finance

As stated earlier, financial issues and part-time job commitments are deeply entwined. (Student A) Student F, who was self-financing, had been working up to twenty hours a week in the first and second year and had had great difficulties balancing this work and study. It had caused the failure of a module in the first year and the transfer to the generic degree route. Student B admitted to having financial problems in every year which really worried him and had affected grades. This became much worse in the final year due to parental financial problems and hence the withdrawal of support. "I think that's why many students pull out because they just cannot get any more money from anywhere." (Student B) When asked if cutting back on socialising had been considered, the reply was "I couldn't do that Keith!", though it was admitted afterwards that this had happened.

Some students had had considerably more parental financial assistance than others, but some parents had actively interfered in their financial affairs. "I filled in the [loan] form and gave it to my Mum to post and she actually changed it without telling me" (Student C), going on to say that living away from home was

preferable even if it meant not being able to afford things. Others were much better off, "I'm lucky enough to get sponsorship". (Student E)

Managing their personal finances appeared to be a great problem. "I spend most of my money at the beginning of the week and then realise that I've no money for food!" (Student E) It was admitted that checking bank statements wasn't enjoyable and caused fear. Student F admitted to being awful at managing money and may have worked more than necessary if finances had been managed better. Although a significant amount of money had been saved during the placement year it completely ran out in the final year and it was necessary to turn to parents in desperation.

Some students were much better at managing their finances. Student D felt that some sort of employment was necessary to supplement the student loan in the first and second years but the loan was completely paid off during the placement year. Student E also managed to pay off much of the overdraft whilst on placement.

These findings from the student interviews will be used in chapter 7 in conjunction with findings from literature in the formation of a questionnaire to be administered to a large numbers of students.

Chapter 7: Devising the Questionnaire

7.1 Introduction

There are two major areas that the questionnaire needs to cover, those related to the student regardless of context and those related to the students' problems or difficulties. For this reason, and bearing in mind the intended use of Data Mining, it was decided that the questionnaire would be divided up into two sections as follows:

1. Demographic Questions
2. Problem Questions

The set of demographic questions are designed to classify and characterise the students whereas the set of problem questions are designed to measure the extent of the problem posed by the question.

7.2 Types of Question

There are two sorts of questions that can be asked in a questionnaire. These are 'open-ended' and 'closed-ended' questions. (Oppenheim, 1992)

7.2.1 Open-ended Questions

"Open-ended questions are used where the issue is complex, where relevant dimensions are not known, and where a process is being explored" (Stacey, 1969)

These are questions where the respondent has a free choice to answer them as they see fit. They offer the respondent more scope in answering the question. They are "more difficult to administer, however, since they require more effort from both the interviewer and the interviewee. Moreover, the qualitative data generated from open-ended questions can be difficult to analyze." (Miller, 2002:9)

7.2.2 Closed-ended Questions

"[C]losed questions should be used where alternative replies are known, are limited in number and are clear-cut." (Miller, 2002:9)

These are questions where the list of alternative responses is known. They are usually referred to as 'multiple choice questions'. They are easier to analyse, quantify, and compare across respondents and are particularly useful when a large number of respondents are expected.

Closed ended questions are easier to analyse and compare respondents. They give the respondents the range of required responses. However, “respondents are forced into what may seem to them an unnatural reply; they have no opportunity to qualify their answers or to explain their opinions more precisely”. (Sheatsley, 1983: 197)

7.2.3 Choice of Question Type

“Free-response questions are often easy to ask, difficult to answer, and still more difficult to analyse.” (Oppenheim, 1992:113)

Because of the large number of anticipated respondents the use of open-ended or free-response questions was excluded. It was decided that all the questions would be ‘closed ended questions’. That is, the respondents are given a fixed set of responses. These questions give the student a fixed set of responses to ensure conformity of answer. When the data analysis is conducted the students can then be grouped in various ways according to their demographics. Commonly established groupings were sought to ensure comparability of data with other such surveys.

This part of the research is aimed at generalisation within this wider research population to find answers to Objective 2, *“Investigate quantitatively, by means of a questionnaire, how widespread these issues are”*.

7.3 Demographic Questions

As mentioned above, it was decided to present the demographic questions to the students with a fixed list of responses thus allowing consistent and manageable results that can be easily analysed. In previous research not much was said about student demographics, though much was said about the diversity of the students. See paragraph 3.4.2 in particular.

Oppenheim (1992) believes that demographic questions should be asked at the end of a questionnaire, “by which time we can hope we have convinced the respondent that the enquiry is genuine.” (Oppenheim, 1992:132) He goes further to suggest that they should always be asked almost apologetically with words such as “now, to help us classify your answers statistically, may I ask you a few questions about yourself ...?” (Oppenheim, 1992:132) He believes that there is a danger in collecting unnecessary socio-economic data. However,

because the demographic data is so crucial to the process of Data Mining it was decided to go against Oppenheim's advice and ask these questions first.

After various investigations and consultations with sources such as student records, Student Services and informal discussions with staff and students a number of questions started to emerge. Some, such as Country of Birth was ruled out in favour of 'Ethnic Group' as the number of possible answers was too great. Other questions needed sensitive wording, such as 'national identity'. Family background was particularly difficult so it was decided to ask about the job of their main family wage earner, so as to avoid issues such as single parent families and mature students.

Care was also taken to ensure that there weren't too many questions and that they would all fit on one page.

Below is the final list of agreed questions:

1. What is your gender?
2. What is the highest post-school education of your parents/guardians?
3. What is your age group?
4. What do you consider your national identity to be?
5. What is your ethnic group?
6. What are your family commitments?
7. What is your home region?
8. What is your position within your family?
9. How many brothers and sisters (including half & step) do you have?
10. Which option best describes the job of your main family wage earner?
11. Which university do you attend?
12. What sort of accommodation do you live in?
13. Do you have a disability?
14. What is your marital status?
15. What is your previous educational background?
16. What is your previous work experience?
17. Do you have work experience relevant to your course?
18. What year of your course are you currently in?

After agreeing the questions the more difficult task of deciding the response options was tackled. Again an extensive discussion and search was done to find standard groupings for the questions. These were used wherever possible. The final list of options can be found in Appendix C.

7.4 Problem Questions

The Problem Questions (student difficulties) are based on the results of the initial student interviews together with the list drawn up from previous research. As above it was decided that they would all be 'closed-ended' or objective questions using a five point scale.

We must first agree on the final categories of questions that need to be asked. Tinto (1993) suggested that there are two sorts of student experiences, academic and social. These will clearly result in some student problems.

Long et al (1995:13) came up with four factors that affect students:

- Institutional factor
- Non-institutional factors
- Student background
- Intra-personal

Yorke (1999:39-46) defined the following factors:

- Poor quality of student experiences
- Inability to cope with the demands of the programme
- Unhappiness with the social environment
- Wrong choice of course
- Matters related to financial need
- Dissatisfaction with aspects of institutional provision

Woodley et al (1987) summarised reasons for withdrawal as:

- Course factors
- Institutional factors
- Study environment factors
- Personal blame
- Motivational factors

When the categories defined by Tinto (1993), Long (1995), Yorke (1999) and Woodley (1987) were considered in conjunction with the authors own observations the following list of categories were finally agreed:

1. Course
2. Institutional
3. Academic
4. Personal
5. Financial

The selection of the questions associated with these categories proved more difficult. Appendix D contains the list of problem areas and suggested questions that were drawn up as the result of previous research modified in the light of the initial student interviews.

The final list of questions to be presented in the questionnaire are as follows:

1. I chose my course in a hurry
2. The course has been too difficult
3. The course is different from what I expected
4. The course has been well taught
5. The study facilities are good (library, computers etc)
6. The support facilities are good
7. I have had no difficulties with student accommodation
8. I have had more failures or referrals than most students
9. At the start of my course I lacked basic skills such as essay writing & note taking
10. My organisational skills are good
11. I do not perform to my best in examinations
12. I have always kept up with my work
13. I have missed more lectures and tutorials than most other students
14. I am satisfied with my level of achievement
15. I dislike criticism from tutors and/or peers
16. I had difficulty settling in when I came to University
17. I have a number of commitments outside University
18. My personal circumstances have changed whilst being at University
19. I have had health problems that have caused me to take time off
20. I have never suffered from stress
21. I have had a number of friendship/relationship problems
22. I lack self confidence
23. My part-time job conflicts with my studying
24. There are too many distractions that affect my ability to study
25. I have had issues with drugs and/or alcohol
26. I give in to peer-group pressure
27. I have always had parental financial support
28. My family's financial circumstances have changed during my time at University
29. I have many financial commitments
30. I am bad at managing my finances
31. I have a large debt apart from my student loan
32. I have sometimes found the course stressful
33. I have considered changing/leaving at some stage
34. My tutors give me support when I need it
35. I have found it easy to make friends at University
36. The induction programme helped me to feel more comfortable at University

A specific problem was the wording of the questions and the slant to put on them. They needed to be as positively worded as possible and not to lead students to giving particular answers. The author decided that all questions would be answered on the same scale. This put a lot of strain on the wording of the questions, but would make them easier to answer by the students. The author also decided to word them as statements rather than questions.

A great deal of effort was applied to reduce the number of questions asked to ensure that the questionnaire would not take too long to fill in and hence not put students off. Thirty-six, though quite extensive was considered suitable for piloting with a small sample of students. The scale used to answer the questions was set at:

1	2	3	4	5
Strongly Disagree	Tend to Disagree	Undecided	Tend to Agree	Strongly Agree

7.5 Administering the Questionnaire

Administering the questionnaire was considered and two main methods were put forward. These were:

- A manual form given out the students and collected in afterwards
- An online questionnaire

The manual form was considered to be easy to formulate and difficult to compile the answers as there was a need to enter all the information into a database afterwards. There was also the problem of non-returns. On the other hand, although the online questionnaire avoided many of the risks of the manual one (the data being entered directly into the database) it would require extensive skills and time to create the facility. Since the author had a previous background in programming it was decided that the time taken to learn the programming language (ASP®) and the web design (Dream Weaver®) was worth the effort. A computerised web-enabled questionnaire was devised. This can be seen in Appendix E.

7.6 Summary

The final questionnaire was piloted by a sample number of students, and with only a few minor adjustments to wording and layout, was made live. SHU students from various years within the ITPA were contacted by email and a number of tutors agreed to allow students to fill them in at the start of a tutorial.

A number of other universities were contacted by establishing, wherever possible a known contact. It was thought necessary to ensure that at least a sample of students from elsewhere should be included in the survey to ensure that the issues found were not specific to SHU.

The results of the questionnaire can then be analysed by quantitative methods using standard statistical techniques and Data Mining tools.

Chapter 8: A Methodology to Use

8.1 Introduction

The questionnaire data was collected over an extended period from April 2005 to January 2006. It was done by means of a fixed response online form and stored in a SQL-Server^{®1} database. SQL-Server[®] was chosen as the database medium because of its versatility. Additional factors were the author's expertise with this database and that it was readily available as a university resource. The database schema can be found in Appendix F. The use of an online entry form with fixed responses ensured conformity of answers and avoided time consuming data entry (see section 7.2.3 Question Type).

Much of the data was collected from SHU; however a sample of data (15.5%) was collected from other similar institutions to ensure that there was a degree of consistency and that SHU wasn't significantly different from other such institutions. A total of 587 questionnaires were collected, 91 of which were from other similar institutions. Although the study was conducted to find issues specifically from computing students at SHU it is expected that many of the finding might find a wider applicability.

Careful consideration of the possible data exploration techniques was undertaken (see Chapter 4 - What's New in this Research). Although Data Mining was chosen as the most effective tool because of its inductive approach, it was decided to pre-process the data by more traditional graphical techniques (Descriptive Data Mining) first in order to become familiar with the data (see Section 9.2(e)).

Data Mining techniques will then be applied in an endeavour to find interrelationships in this data. **Data Mining** is a technique used to find patterns within data and examine inter-relationships within the data; it looks for trends or anomalies without knowledge of the meaning of the data. (Hyperdictionary, 2004) This will help to provide enlightenment for Objective 3, *"Explore, by means of Data Mining techniques, the interrelationships between the issues"*.

Care will need to taken at this stage to ensure that the analysis remains focused on the central aim of the study, "to explore the issues that affect ... students and contribute to dropping out, with the view to possible tutor intervention to improve

¹ SQL-Server[®] is Microsoft's Relational Database

retention rates” (see Section 2.1). It is expected that the data might reveal other avenues for investigation such as gender and cultural issues which are beyond the scope of this study but might constitute a further study at a later stage.

Like all good projects there is a need for a methodology to form a template to work to. There are two well established Data Mining Methodologies, SEMMA and CRISP-DM. SEMMA (Sample, Explore, Modify, Model, Assess) was devised by the SAS Institute® (see Section 4.2.4). It forms a rigid process to be followed when using the SAS Data Mining software Enterprise Miner®. However since the author chose to use the SPSS® software Clementine® (see Section 4.2.4), it was decided that it was more appropriate to use the methodology developed and recommended by them. This methodology is called CRISP-DM (Cross-Industry Standard Process for Data Mining). It is outlined in the next section.

8.2 CRISP-DM – A Methodology

8.2.1 History

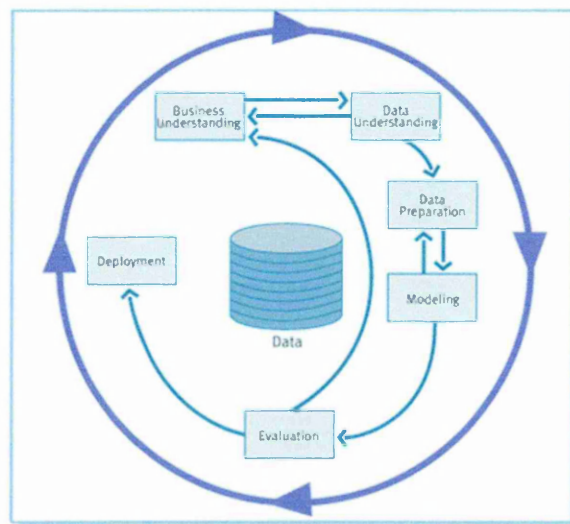
CRISP-DM was conceived in 1996 by organisations that were pioneering Data Mining. These were SPSS® (then ISL) who had produced their Clementine® software just two years previously, NCR®, as an attempt to deliver added value to its Teradata® Data Warehousing customers and DaimlerChrysler® (then Daimler-Benz®) who were considered to be ahead of the game as regards Data Mining in an industrial context. At that time market interest in Data Mining was becoming intense and there were signs of a widespread uptake. Up to this point organisations had developed their own approaches to Data Mining as they went along. None were sure that they were doing it right.

There was a clear desire that the experiences and expertise of these early data miners should be shared with a wider community. They described it as “both exciting and terrifying”. (CRISP-DM, 2000:3) They wanted to produce a standard process model that was not tied to a particular software product (non-proprietary) and was freely available to all practitioners. The focus is on business issues as well as technical analysis and is intended to be a framework for guidance.

The project was funded by the European Commission and drew upon a number of other Data Mining practitioners and vendors. This was done through the CRISP-DM Special Interest Group (SIG).

8.2.2 Reference Model

CRISP-DM divides the Data Mining lifecycle into six phases, though it is not intended that the sequence of these phases are to be adhered to strictly. It is expected that the developer(s) will move backwards and forwards through the model as required depending on the outcome of a particular phase. Each phase has a number of defined tasks. The arrows on the diagram below indicate the “most important and frequent dependencies between phases.” It is important to note that even when the project has been deployed “subsequent data mining processes will benefit from the experiences of previous ones.” (CRISP-DM, 2000:13)



Phases of the CRISP-DM Reference Model (CRISP-DM, 2000:13)

8.2.3 The Stages of the Reference Model

As stated above, there are six defined stages in the process model, each of them well defined and clearly set out with defined deliverables expected at every stage. It might be said that there is an element of overkill in this methodology, and perhaps there is. However, it does force the analyst(s) to look at every aspect of the process to ensure that everything is visited. This section relies heavily on the *CRISP-DM Step-by-step Data Mining Guide* (CRISP-DM, 2000).

1. Business Understanding

The initial stage focuses on understanding the project objectives from the perspective of the business. It then converts this knowledge into data mining terms and sets up a preliminary plan in order to realise these objectives, giving success criteria where appropriate.

a) Determine Business Objectives

Here the analyst must examine the background of the project, set the business objectives and possible success factors.

b) Access Situation

At this point the necessary requirements, assumptions and constraints are set out and the necessary resources are defined. Risks and contingencies are examined and the costs and benefits analysed.

c) Determine Data Mining Goals

Next comes the setting of the Data Mining goals together with their respective success criteria.

d) Produce Project Plan

Finally the project plan is drawn up and an initial assessment of tools and techniques required is made.

2. Data Understanding

This stage starts with the initial data collection and attempts to familiarise the analysts with the data, identifying quality problems, gaining insights into the data where patterns might be found.

a) Collect Initial Data

This part speaks for itself. Data are collected from source systems, such as point-of-sale, or other such transactional systems. It must be pointed out that there might be a number of source systems that the data is collected from where the data might be in significantly different formats. This is not a trivial task. An 'Initial Collection Report' will be produced.

b) Describe Data

Once the required data has been identified and the format defined then a 'Data Description Report' is produced. This will include a list of the variables, their data types and what use is likely to be made of them.

c) Export Data

Having identified the required data and defined its format it is then exported to a medium that is dedicated to the purpose of data cleansing, formatting and merging. The area is usually termed as the 'Staging Area'. It is likely to be a relational database, though some vendors now offer specific tools for this purpose. They are called ETL tools because

they allow the analyst to Extract, Transform and Load the data. 'Extract' it from the source systems, 'Transform' it into the required format and 'Load' it into the Data Warehousing or Data Mining environment. This is the Data Mining 'Extract' stage.

At the end of this process a 'Data Exportation Report' is produced.

d) Verify Data Quality

Once the data has been exported from the source systems it is examined for data quality issues. These might be things such as:

- Incomplete or missing values
- Corrupted values
- Out of range values
- Wrong data
- Duplicate data

This is part of the 'Transformation' phase. At the end of this process a 'Data Quality Report' is produced.

3. Data Preparation

This phase covers all the activities required to transform the data and construct the Data Mining dataset from the source system that originally captured the data. There might be many heterogeneous (different) data sources, so this process is far from trivial.

a) Select Data

Using the previously created reports the required data is examined and the final format is agreed. A rationale for inclusion or exclusion of certain data items is determined.

b) Clean Data

The data is then cleaned with the use of the 'Data Quality Report' as a basis. If data has come from a number of different source systems it will be necessary to clean each part of the data separately. However it might not be possible to clean all the data as some of the required information might not be available. Such things as default values and data exclusion might be needed. At the end of the process a 'Data cleansing report' is produced.

¹ Care must be taken here, the use of the word 'Extract' in the ETL process is different from that in CRISP-DM

c) *Construct Data*

The data components are now defined and their respective attributes established. Not all attributes that exist in the original source data are going to be preserved. It is important to point out that if data has been captured from different source systems only data that is common to the different systems can be used. It is not usually possible to 'invent' data that has not been recorded earlier. Individual records are now generated.

d) *Integrate Data*

If the data is coming from more than one source then it will need to be merged at this point into an integrated whole. This is the final stage in the 'Transformation' part of the ETL process.

e) *Format Data*

Most Data Mining software allows the data to be imported in various formats. However they all require it to be in a 'flat-file' format. That is, the data needs to be in one table in a clearly defined format.

4. Modelling

The modelling techniques are now selected and a number of trial runs are undertaken to define the parameters and establish their optimum values. This is likely to require looping back to the data preparation phase to adjust the data set. Each stage requires documentation.

a) *Select Modelling Technique*

The modelling techniques (e.g. Decision Trees or Neural Networks) are reviewed and the most appropriate ones are selected. Some assumptions might have to be made at this stage. For instance, no missing values are allowed or that all the variables have a uniform distribution.

b) *Generate Test Design*

Before the models are created it is necessary to establish a procedure to test the model's quality and validity. For instance it is usual to use part of the dataset to train the model and another part to test it. In this way an estimate of its reliability can be established.

c) Build Model

The modelling tools are now run on the prepared data. They might be run several times in order to adjust the parameters. The models are described and a report prepared. Any difficulties encountered will be documented.

d) Assess Model

The analyst now interprets the models in the light of his/her business knowledge, the pre-established success criteria and test design. Consultation with the business analysts and other business experts will be needed here. The models might be 'tuned' by adjusting their parameter settings. These changes are documented.

5. Evaluation

The models now appear to conform to expectation from a data analysis perspective. However, before they are deployed it is necessary to evaluate them to ensure they achieve the business objectives. "A key objective is to determine if there is some important business issue that has not been sufficiently considered." (CRISP-DM, 2000:14)

a) Evaluate Results

This step assesses whether the model(s) meet the business objectives. If time and budget permits the models might be piloted in the real environment. If all is well then the models are then given the 'seal of approval'.

b) Review Process

The model should now satisfy the business needs. It is therefore important to review the whole process thoroughly to ensure that nothing has been overlooked. This also covers quality assurance, "did we correctly build the model? Did we only use attributes that we are allowed to use and that are available for future analyses?" (CRISP-DM, 2000:31)

c) Determine Next Steps

Finally the assessment results and process review are examined and a decision is made as to whether to proceed to deployment or to make additional iterations. Possible action scenarios are listed and a final decision is made and documented along with a rationale.

6. Deployment

Creating the model(s) is not the end of the story. "Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it." (CRISP-DM, 2000:14) It is likely that it is the customer and not the analyst that carries out the deployment.

a) Plan Deployment

The evaluation results are examined and a deployment strategy is formulated. The following activities might be undertaken:

- " – Summarize deployable results
 - Develop and evaluate alternative plans for deployment
 - Decide for each distinct knowledge or information result
 - How will the knowledge or information be propagated to its users?
 - How will the use of the result be monitored or its benefits measured
 - Decide for each deployable model or software result
 - How will the model result be deployed within the organization's systems
 - How will its use be monitored and its benefits measured?
 - Identify possible problems when deploying the data mining results"
- (CRISP-DM, 2000:60)

b) Plan Monitoring and Maintenance

A detailed plan needs to be devised that takes account of the specific type of deployment. This is particularly important as Data Mining becomes part of the day-to-day processes of a business. It should attempt to answer some or all of the following tasks:

- What things could change?
 - How will accuracy be monitored?
 - When should the business stop using the model?
 - What things could change?
 - What should happen if the model can no longer be used?
 - Will the business objectives change over time?
- (CRISP-DM, 2000:61)

c) Produce Final Report

A project report is written which should include how well the initial goals were met.

d) Review Project Experience

Assess the successes and failures of the project and what could be done to improve the situation. This might include any blind alleys that were pursued and any pitfalls that were encountered. "Interview all significant people involved in the project and ask them about their experiences during the project." (CRISP-DM, 2000:62)

Chapter 9: Mining the Data

9.1 Business Understanding

"We are drowning in information but starved for knowledge." (Naisbitt, 1982)

In the case of this project the 'business' may be defined as the Educational Process, the specific part of the business being 'Student Retention'. The preceding chapters of this study might be interpreted as gaining a 'Business Understanding'. The author has worked in various forms of education for more than thirty years, gaining an insight into the educational process and the problems that face students. The author believes that the literature review and the student interviews have equipped him with knowledge of the fundamentals of 'student retention' and a thirst to find out more.

In many ways the students and the academic/administrative staff appear to occupy different, parallel worlds that seem to lack understanding of each other. There is a definite need for increased interaction.

a) Determine Business Objectives

The project had six objectives that it set out to achieve (see Section 2.2).

The one that is applicable here is Objective 3 shown below:

Objective 3

Explore, by means of Data Mining techniques, the interrelationships between the issues.

Translating this into business terms we might arrive at:

1. Determine interrelationships in the data
2. Establish the factors that contribute to student attrition
3. Establish demographic profiles of students most at-risk of attrition

These are set to achieve the aim of 'increasing student retention rates'.

Having defined the Business Objectives it is now important to define the success criterion:

Give useful insights into the relationships between the student demographics and the associated student problems encountered.

This might be deemed to be rather subjective, however it will be the job of the Focus Group to assess this in the light of the findings and formulate the recommendations.

b) *Access Situation*

There is a need for clean data produced from SHU and a number of similar institutions. The amount of data required is difficult to assess. John and Langley (1996) experimented with sample sizes ranging from 300 to 2,180 and found that they all fell within a 2 percent range of accuracy as measured by the confidence factor in the data mining model (see Section 9.4). Oates and Jensen (1998:254) note "Increasing the amount of data used to build a model often results in ... no significant increase in model accuracy." It was decided therefore obtain at least 500 rows of data. In fact 587 were collected. With less than this there is a likelihood of insufficient evidence being available. However, no top limit is necessary provided that all the data is drawn from the same population.

The data was originally collected around the middle of the second semester of the academic year 2004-5. However there were insufficient responses at this time and the author had to wait until the later part of the first semester of the academic year 2005-6 in order to top up the data responses. There was a risk here that the second sample would differ in profile to the initial sample. Great care was taken, to ensure that this was not the case. However, when referring to the problems of data availability Berry and Linoff (2000) state:

"Often the right data is simply whatever data is available, reasonably clean, and accessible." (Berry & Linoff, 2000:49)

c) *Determine Data Mining Goals*

Translating the business objectives outlined above into Data Mining goals is a tricky one. However after careful consideration the following were derived:

1. Determine interrelationships within the response data
2. Predict the likelihood that a student will leave given other known factors
3. Determine clusters of students with similar response patterns and ascertain which cluster(s) are most at risk of attrition

d) *Produce Project Plan*

As set out in Section 4.2 Data Mining falls into two distinct areas, supervised learning and unsupervised learning. It is intended to use both types of learning to achieve the goals set above.

Goal 1

The first goal: '*Determine interrelationships within the response data*' specifies no output variable and hence appears to be a straightforward unsupervised activity. The most likely contending model is 'Rule

Association'. Either the GRI model or the Apriori model can be used (see Section 4.2.5 above). They are both specifically designed to meet such a need. The outcome of these models could be taken further by applying 'Market Basket Analysis' to discover the principal issues that are associated with these outcomes.

Goal 2

The second goal: *'Predict the likelihood that a student will leave given other known factors'* presents a fairly standard Data Mining task. It is clearly supervised learning as there is a known (or given) output 'leave'. (See Section 4.2.5 above) Point 33 of the problem section of the questionnaire asks the students to respond to the following statement: "I have considered changing/leaving at some stage" (see Appendix D). This means a model can be built using the demographic responses and other problem responses as inputs. Clearly some work will be needed to reduce the inputs to the most significant ones (variable reduction).

Two models appear to fit the bill, Neural Networks and Decision Trees. However, Neural Networks work better with a significant amount of data. (See Section 4.2.5) The author felt that 587 rows were probably below a desirable amount but decided to test it out alongside a Decision Tree model.

Goal 3

The third goal: *'Determine clusters of students with similar response patterns and ascertain which cluster(s) are most at risk of attrition'* is an ideal goal to be realised using a clustering modelling tool such as Kohonen Networks or K-Mean. On balance there is little to choose from between them so the author selected Kohonen Networks as he has had more experience with it.

It will be necessary to experiment with various numbers of clusters until stable and meaningful results are established. A great deal of trial and error is expected here. Kohonen Networks will establish the clusters but it won't determine the nature of the components of these clusters. Tedious methodical analysis will be needed to discover these.

Post-script

Once the initial investigation has been conducted the results from one model might well be used within another model, hence moving the analysis forward iteratively.

9.2 Data Understanding

“The secret of success is to know something that nobody else knows.”
(Aristotle Onassis, n.d.)

This project is somewhat unique in that the source database has been specifically designed to collect the required data. There is only one database and all the responses are selected from a given list. It has to be assumed that the data is clean as all the responses are anonymous and hence there is no way of correcting it.

a) *Collect Initial Data*

The data was collected via an online input form as shown in Appendix E. This form was in two parts. Part 1 contained demographic questions with fixed answers chosen from combo boxes. Part 2 dealt with the problems the student may have encountered during their time at university. The data was stored in a single relational database located on a Microsoft SQL-Server® platform. The database schema can be found in Appendix F. This database has full referential integrity¹. Because the data is collected in the required format then it was felt that an 'Initial Collection Report' was unnecessary and the schema given in Appendix F was used as a substitute.

b) *Describe Data*

A list of all variables, together with their data types can be found in Appendix G. To aid efficiency of data analysis it was decided to give the variables meaningful names. Appendix G can be considered to be the 'Data Description Report'. All variables will initially be considered but variable reduction is expected to occur during the Data Mining process.

c) *Export Data*

The database was copied in its entirety, to a parallel database. This was done so that the data could be manipulated without contaminating the original source data. Should any problems occur in the cleansing process then it would be possible to start again with the original source data. This database acted as a staging area for the data, that is, a temporary storage area where the data is processed before being made available for Data Mining. A 'Data Exportation Report' is not necessary as the exported data is in the same form as the source data.

¹ Referential Integrity means that no data can be inserted into the transaction table without associated records being found in the related tables

d) Verify Data Quality

The data was then examined for quality issues. The input data was transacted, that is it was only stored in the database once all the questions had been answered. However, since the first screen asked for demographic responses with defaults, it was technically possible for a student to accept all these defaults and move straight on to the next screen. The data was checked using SQL queries to ensure that there was not a preponderance of default records in the database; however none were found.

The data was then examined for specific errors as follows:

1. Incomplete or missing values
2. Corrupted values
3. Out of range values
4. Wrong data
5. Duplicate data

Because of the specific method of collecting the data it was not possible to have incomplete or missing value errors, nor was it possible to have out of range values. However the other types of errors were possible.

The data was checked by the use of SQL queries. All values were checked for corruption and none was found. Wrong data is much more difficult to check for. Since the data was anonymously collected there was no way that a student response could be checked against an expected response. Wrong data would be the product of a mistake being made by a student on entry of the data. This could be accidental or deliberate. Since the questionnaire was administered anonymously with little or no chance of correction, an assumption had to be made that there were few of these errors.

e) Pre-processing the Data

This section has been added by the author as a pre-cursor to the modelling that will be done. Exploring the data to find proportions of respondents with different responses will aid greatly the data understanding. This might be described as Descriptive Data Mining.

NOTE: Though most British students stated their nationality to be British there were some that considered themselves to English, Scottish, Welsh or Northern Irish.

Demographic Proportions

	%	Number
Female	19.59	115
Male	80.41	472

Gender split

	%	Number
College	28.62	168
None	24.7	145
Other	8.52	50
University	38.16	224

Post-school education of parents/guardians

	%	Number
Under 21	34.92	205
21 to 25	56.22	330
26 to 30	4.6	27
31 to 40	3.24	19
Over 40	1.02	6

Age group profile

	%	Number
British	76.49	449
English	17.72	104
Irish	0.17	1
Northern Irish	0.17	1
Other	4.09	24
Scottish	0.17	1
Welsh	1.19	7

Nationality

	%	Number
African	1.19	7
Bangladeshi	0.85	5
Caribbean	0.34	2
Chinese	1.7	10
Indian	7.67	45
Other	3.07	18
Pakistani	6.13	36
White	77.85	457
White and Black African	0.51	3
White and Black Caribbean	0.68	4

Ethnic groups

	%	Number
Dependent children	5.62	33
Elderly or sick parent(s)	3.24	19
None	85.69	503
Other	4.94	29
Sick partner	0.51	3

Family commitments

	%	Number
East	2.73	16
East Midlands	15.84	93
London	1.02	6
North East	2.39	14
North West	12.95	76
Northern Ireland	0.34	2
Other	2.9	17
South East	10.05	59
South West	3.58	21
Wales	2.21	13
West Midlands	6.81	40
Yorkshire & Humber	39.18	230

Home Region

	%	Number
First or only child	49.4	290
Second child	30.83	181
Third child	12.78	75
Fourth child	3.41	20
Greater than fourth child	3.58	21

Position in Family

	%	Number
None	10.9	64
One	41.57	244
Two	24.7	145
Three	10.9	64
More than three	11.93	70

Numbers of brothers/sisters

	%	Number
Unskilled	6.3	37
Partly skilled	4.94	29
Skilled manual	17.21	101
Skilled non-manual	4.6	27
Technical	7.16	42
Managerial	20.78	122
Professional	39.01	229

The job of the main family wage earner

	%	Number
SHU	84.5	496
Other	15.5	91

University Attended

	%	Number
Halls of residence	20.1	118
Other	0.68	4
Parental home	23.68	139
Private owned	5.45	32
Rented (furnished)	45.83	269
Rented (unfurnished)	4.26	25

Accommodation type

	%	Number
No	89.95	528
Yes	10.05	59

Disability

	%	Number
Divorced	0.51	3
Long-term partner	11.75	69
Married	3.07	18
Single	84.67	497

Marital Status

	%	Number
School only	24.7	145
College	66.27	389
University	6.98	41
Polytechnic	1.02	6
Other	1.02	6

Previous Educational background

	%	Number
None	6.64	39
Part-time only	41.57	244
Under 1 year full-time	13.29	78
1 to 2 years full-time	27.09	159
More than 2 years full-time	11.41	67

Work experience

	%	Number
No	43.95	258
Unsure	6.3	37
Yes	49.74	292

Experience relevant to course

	%	Number
First	19.59	115
Second	21.64	127
Third	20.1	118
Fourth	36.63	215
Over fourth	2.04	12

Year of study

On close examination of the above charts it can be seen that some of the demographic splits are likely to be of minimal use. The nationality reveals only twenty-five students that are Non-British (if we put together the component parts of Britain).

The range of ethnic groups is rather disappointing with almost 80% being white. Though, this is roughly what would be expected, the non-white are very widely spread and are likely to be insignificant when attempting to draw conclusions from them.

Home region reveals a high proportion of the students coming from the region in which the university is situated (Yorkshire & Humber) and many of the rest coming from the neighbouring regions of the East Midlands and the North West. There may be scope to regroup these into 'University Catchment' and 'Other'.

Looking at the position in family, it might have been wiser to have had the highest grouping as 'Greater than third child' since the proportions of 'Fourth child' (3.41%) and 'Greater than fourth child' (3.58%) are very small. Putting these two together would give a more significant proportion (6.99%). However, when examined in conjunction with the number of brothers and sisters there appears to be something of a discrepancy with 11.93% stating that they have more than three brothers and sisters. This is probably accounted for when putting step brothers and sisters into the picture.

A cause for concern, when looking at the data from a 'Widening participation' perspective, is the small number of students coming from the unskilled and partly-skilled family backgrounds. These only account for 11.24% of the student profile. Even when skilled-manual workers are added to this it only accounts for 28.45%. More research here would be useful.

Disability and university attended have already been grouped from the original data as the proportions of the different groups were far too small to be of meaningful value.

<u>Problem Proportions</u>		<u>Number</u>		<u>Percentage</u>	
QUESTION		No	Yes	No%	Yes%
1. Hurried choice of course		437	150	74%	26%
2. Difficult course		472	115	80%	20%
3. Course different from expectation		272	315	46%	54%
4. Taught Badly		451	136	77%	23%
5. Facilities unsatisfactory		536	51	91%	9%
6. Support unsatisfactory		508	79	87%	13%
7. Accommodation difficulties		463	124	79%	21%
8. Failures more than most		452	135	77%	23%
9. Skill deficiency at start of course		455	132	78%	22%
10. Organisation skills poor		460	127	78%	22%
11. Examination performance below expectations		279	308	48%	52%
12. Behind with work		439	148	75%	25%
13. Missed more lectures and tutorials than most		485	102	83%	17%
14. Achievement below expectation		300	287	51%	49%
15. Criticism disliked		473	114	81%	19%
16. Settling in was difficult		461	126	79%	21%
17. Outside commitments high		306	281	52%	48%
18. Personal circumstances changed		385	202	66%	34%
19. Time off with health problems		490	97	83%	17%
20. Stress problems		256	331	44%	56%
21. Friendship or relationship problems		446	141	76%	24%
22. Lack self confidence		397	190	68%	32%
23. Part-time job conflicts with studies		455	132	78%	22%
24. Too many distractions that affect ability to study		342	245	58%	42%
25. Issues with drugs and/or alcohol		517	70	88%	12%
26. Give in to peer-group pressure		489	98	83%	17%
27. Parental financial support lacking		395	192	67%	33%
28. Family's financial circumstances changed		433	154	74%	26%
29. Financial commitments high		392	195	67%	33%
30. Bad at managing finances		411	176	70%	30%
31. Large debt apart from my student loan		445	142	76%	24%
32. Have sometimes found course stressful		205	382	35%	65%
33. Considered changing/leaving at some stage		325	262	55%	45%
34. Tutor support sometimes lacking		448	139	76%	24%
35. Found it difficult to make friends at University		511	76	87%	13%
36. The induction didn't help to feel more comfortable at University		412	175	70%	30%

Figure 1

9.3 Data Preparation

As stated in Section 8.2.3 above, this phase covers all the activities required to transform the data and construct the Data Mining dataset from the source system that originally captured the data. There is only one data source, so this process is relatively straight forward.

a) Select Data

All data records and attributes were kept as they were specifically stored for the purpose of data mining.

b) Clean Data

All the data was considered to be clean so again nothing was needed to be done at this stage.

c) Construct Data

The problem here was how to convert the five point scale into variables that were more appropriate for data mining purposes. A five point scale was used to allow the respondents freedom to show their level of response. However, for the purpose of data mining, all that is needed is whether the respondent felt that there was an issue or not. All the scales then were converted into Yes/No responses. In addition, because the same scale had been used for all responses, some of the responses needed rewording to allow the 'Yes' response to be the problem. For instance Statement 14 originally said 'I am satisfied with my level of achievement'. To ensure that this question gave a 'Yes' response to the problem of achievement it was reworded in the Data Mining dataset as 'Achievement below expectation'. Hence care needed to be taken to ensure that the responses were correctly coded as Yes/No. Responses of 1 and 2 were grouped together as were responses of 4 and 5.

Where to put the response of 3 (undecided) presented quite a dilemma. For instance, in a response of 3 to the statement 'I have a large debt apart from my student loan' might indicate that the student wasn't bothering to check his/her debt level and might indeed have a large debt. However, so as not to over-estimate the problems it was decided that ALL responses of 3 should be placed in the 'No' category. The conversion criteria can be found in Appendix H.

There was a danger here however, that the data would lose some of its richness. Most of the statements were originally worded to positively reflect the problem that was being measured. However, those in red were reversed so that they were in the right format for Data Mining (see Figure 1 above).

d) Integrate Data

This stage was omitted since there was only one data source.

e) Format Data

Since the data had been stored in a relational database it needed to be converted into a format that could be accepted by Clementine®, the chosen Data Mining software. The acceptable format of Clementine® is a single CSV file with a tab, comma or space delimiter. Comma delimiters were chosen as it was the default for SQL-Server®. If you examine the database in Appendix F, the original source database, you can see that it is made up of a number of related tables. The data had to be flattened into the single flat file format of Appendix I.

Below is a sample of the data in the format of Appendix I. The actual data has too many columns to meaningfully display on screen, so some of the columns have been removed for display purposes.

The first row of the dataset would normally be the variable (column) headings, however this has been omitted to allow the data to be displayed meaningfully.

```
Male,21 to 25,Bangladeshi,None,North West,Unskilled,No,Third,Y,Y,Y,Y,N,N
Male,21 to 25,White,Dependent children,Yorkshire & Humber,Managerial,No,Third,N,N,Y,N,N,N
Male,21 to 25,White,None,North West,Professional,Yes,Third,N,N,Y,Y,N,N
Male,Under 21,Pakistani,Other,Yorkshire & Humber,Professional,No,Third,N,N,N,N,Y,N,N
Male,Under 21,White,None,East Midlands,Managerial,No,Third,N,N,N,Y,N,N
Male,21 to 25,White,None,London,Professional,Yes,Third,N,N,Y,Y,N,N
Male,26 to 30,Other,Dependent children,Yorkshire & Humber,Professional,Yes,Fourth,N,N,Y,Y,N,N
Male,Under 21,Chinese,None,Yorkshire & Humber,Professional,Unsure,Third,N,N,Y,Y,N,N
Female,21 to 25,White,None,Yorkshire & Humber,Technical,Yes,Third,N,N,N,N,N,N
Male,21 to 25,Pakistani,None,Yorkshire & Humber,Professional,No,Third,N,N,N,Y,N,N
Male,21 to 25,White,None,Yorkshire & Humber,Unskilled,No,Third,Y,Y,Y,Y,N,N
Male,Under 21,Pakistani,None,Yorkshire & Humber,Partly skilled,Unsure,Third,N,N,N,Y,N,N
Male,21 to 25,Indian,None,East Midlands,Unskilled,No,Third,Y,N,Y,N,N,N
Male,21 to 25,Indian,Other,East Midlands,Skilled manual,Unsure,Fourth,Y,N,Y,Y,N,N
```

9.4 Modelling

The modelling techniques suggested in 9.1(d) are now examined and final decisions are made. In section 9.1(c) three Data Mining goals were set. These are reproduced below for convenience:

1. Determine interrelationships within the response data
2. Predict the likelihood that a student will leave given other known factors
3. Determine clusters of students with similar response patterns and ascertain which cluster(s) are most at risk of attrition

A whole variety of modelling tools are available in the SPSS Clementine[®], the chosen software. Each of the goals above was examined and the most appropriate tool selected. In some cases there was a choice of suitable tools. The actual tool used was chosen on its merits in the specific situation it was to be used.

Only two adjustments were made to the dataset. The first was to group students into 'SHU' and 'Other' rather than showing the different universities. This was, firstly to keep the anonymity of the data, and secondly many institutions gave a small number of responses and would have been insignificant in the bigger picture. The second adjustment was to group all the disabled students together as the individual disabilities were insignificant. However, this could be looked upon as divisive and hence little or no use will be made of the output produced.

In the sections that follow the application of the three rules are documented.

9.4.1 Rule 1

Determine interrelationships within the response data

a) *Select Modelling Technique*

The wording of this goal suggests that a non-supervised form of learning might be most appropriate since no output has been determined except for “interrelationships”. Rule Association appeared to be the starting position as it has been developed especially for this kind of analysis.

“Association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules show attribute value conditions that occur frequently together in a given dataset.” (Bertino et al, 2004:5)

There are two different Rule Association modelling tools in Clementine®; these are **GRI** (Generalised Rule Induction) and **Apriori**. Both these techniques are equally reliable so it is immaterial which one to use. Both modelling tools were tested and the author chose **Apriori** as it appeared to home in to an establish pattern of relationship faster.

b) *Generate Test Design*

There are a number of parameters associated with Rule Association models. The first is ‘Minimum rule support’ (Support). This is the percentage of data-points that backup this relationship. A value of around 25% is a good starting point. The second parameter is ‘Minimum support confidence’ (Confidence). This is how confident we are that the established rule is correct. We will start high at 85% and see if any relationships are evident and gradually lower this until a manageable number are found that are still acceptable, setting the lowest acceptable confidence at 70%. Only rows of data that show true against all the variables will be considered. The consequent behaviour is determined by its antecedents.

c) *Build Model*

The Apriori modelling tool was now run on the prepared data. Several runs of the model were made to optimise the parameter settings. Only three have been documented below. The first one shows the model with the initial parameters of Support 25% and Confidence 85%.

FACTOR 1	FACTOR 2	FACTOR 3
Course	Stress	Leave

Support = 25%, Confidence = 85%

This revealed only one rule which shows a strong relationship between finding the 'course stressful', suffering 'general stress' and considering leaving. Or putting it in terms of student attrition 'course stressful' and 'general stress' are likely to cause the student to consider leaving or at least change course. This doesn't appear to be very startling until you examine the evidence. 25% (support) of the data collected showed this relationship and the modelling tool was 85% (confidence) certain that the relationship was true for the population in general.

This is a useful starting point, but one rule is insufficient, so the parameters were gradually reduced to see the effects. At a Support of 25% and Confidence of 80% still only two rules were produced.

FACTOR 1	FACTOR 2	FACTOR 3
Course	Stress	Leave
Course	Stress	Examinations

Support = 25%, Confidence = 80%

After a number of adjustments of the model the parameters were finally set at a Support of 25% and Confidence of 70%. This delivered 22 rules which seem somewhat manageable whilst still at or above the preset minimum confidence level of 70% and minimum support of 25%.

FACTOR 1	FACTOR 2	FACTOR 3
Course	Stress	Leave
Course	Stress	Examination
Course	Examination	Leave
Course	Stress	Different
Course	Confidence	
Course	Stress	Commitments
Course	Different	Distraction
Course	Work	
Course	Different	Examinations
Course	Examinations	Commitments
Course	Different	Leave
Course	Examinations	Distractions
Course	Stress	
Course	Leave	
Course	Distractions	
Course	Stress	Confidence
Course	Circumstances	
Course	Financial	
Course	Different	Commitments
Course	Examinations	
Course	Commitments	
Course	Management	

Support = 25%, Confidence = 70%

The above table shows the relationship between the variables in descending order of significance.

Having now isolated the most significant student problems we now move on to complete the process of 'Market Basket Analysis' (MBA). That is, set the output to the principal group of problems and run a Decision Tree modelling tool on the dataset.

Analysing the output above the most significant problems appear to be:

- Too many distraction that affect ability to study (Distraction)
- Stress problems (Stress)
- Sometimes found the course stressful (Course)
- Examination performance below expectations (Examinations)

A new variable 'Leave_Questions' was created using the following rule:

```

IF      Distractions = "Y" AND
        Stress = "Y" AND
        Course = "Y" AND
        Examinations = "Y"
THEN   Leave_Questions = "True"
ELSE   Leave_Questions = "False"

```

The student problem of 'Considered leaving or changing' (Leave) has been omitted as it is considered to be a product of other considered issues and the focus of the research.

Using this new variable 'Leave_Questions' as an output, a Decision Tree (C5.0) modelling tool was run. This created a new variable \$C-Leave_Questions which is a calculated field that measures the accuracy of the built model. C5.0 takes half the data set to train the model and the second half of the data set to test it. It then checks the accuracy of the model by measuring the predicted value (established by the model) against the actual value recorded in the data.

d) Assess Model

It is testing the output variable Leave_Questions against the calculated variable \$C-Leave_Questions that shows the accuracy of the model. This is set out in the matrix below:

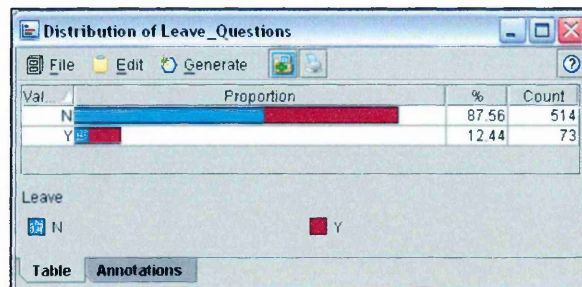
\$C-Leave_Questions			
Leave Questions		N	Y
N	Count	360	154
	Row %	70.039	29.961
Y	Count	23	50
	Row %	31.507	68.493

Though a better accuracy might have been hoped for, just taking the four student problems above still gives a predicted accuracy of 68.5% (68.493), which considering the size of the data sample (587) means that we would get it right more than twice out of three times, certainly significantly better

than chance. Getting it right means the student response being 'Y' to all the questions.

Having now established the accuracy of the model, the object of the exercise 'to predict' where combination of responses are good predictors of 'Leaving or changing' we need to look at 'Leave' against 'Leave_Questions'. Graphically this is shown below.

The bars represent the Leave_Questions and the colour represents the answer to the 'Leave' statement. Although the model is far from perfect it does show that the positive prediction is relatively high. However, it gets it wrong a lot too, though it is still significantly better than chance.



It can be seen from the following matrix, that the model is 68.5% (68.493) accurate, the same as the Decision Tree model!

		Leave	
Leave_Questions		N	Y
N	Count	302	212
	Row %	58.755	41.245
Y	Count	23	50
	Row %	31.507	68.493

It is also of interest to see if the predicted \$C-Leave_Questions is a good measure of 'Leave'.

		Leave	
\$C-Leave_Questions		N	Y
N	Count	320	233
	Row %	57.866	42.134
Y	Count	5	29
	Row %	14.706	85.294

This shows an even better success rate than the leave questions themselves, which seems extraordinary! (85.3% against 68.5%)

Summary

The rule was investigated by Rule Association and the variable relationships were found (see table on page 83 above). These were established at a confidence level of 70% and support of 25%. The table below lists the most significant problems. Those rated high were the problems that were established first at a higher level of confidence.

Rate	Problem
High	Have sometimes found course stressful (Course)
High	Stress problems (Stress)
	Course different from expectation (Different)
	Lack self confidence (Confidence)
	Outside commitments high (Commitments)
High	Too many distractions that affect ability to study (Distractions)
	Behind with work (Work)
High	Examination performance below expectations (Examinations)
	Personal circumstances changed (Circumstances)
	Bad at managing finances (Management)
High	Considered changing/leaving at some stage (Leave)

9.4.2 Rule 2

Predict the likelihood that a student will leave given other known factors

a) Select Modelling Technique

A great deal of data, both demographic and problem orientated, was collected about each of the students. We need to look at the relationship between these and how they can be used effectively to predict whether a student will leave or not. It must be explained that collecting data about students that have left is extremely difficult so to measure this the author has decided to use the student response to statement 33, 'I have considered changing or leaving at some stage' as an substitute for this. Clearly this can only be an estimate, but it does reflect the type of students that are likely to leave, but will be an overestimate and hence any model built using this as an output will reflect this overestimate.

Since the above goal is looking for a specific output, 'Leave', then it requires a supervised modelling tool. A Decision Tree modelling tool such as C5.0 is a good contender for this. We could use a Neural Network, but because of the size of the sample it is likely to give a poor level of accuracy. (See paragraph 9.4.2(d))

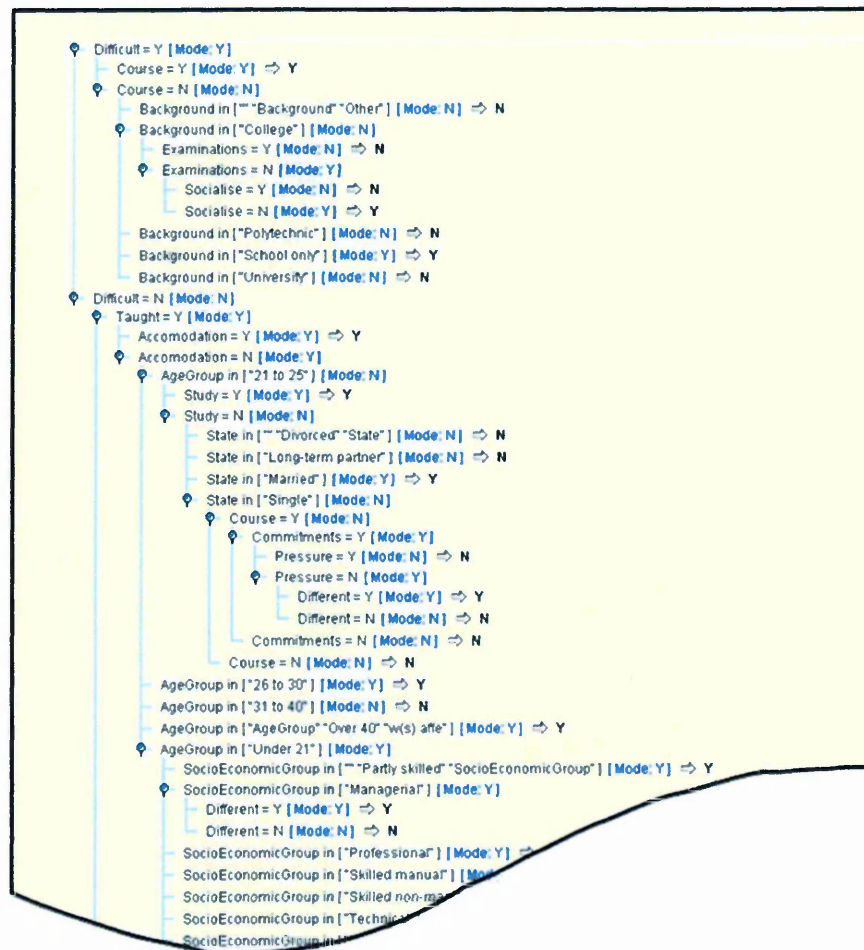
b) Generate Test Design

We will use the in-built mechanism for testing as we did with the last model. A high level of accuracy (80%) will be attempted first, if this is not attainable then the parameters of the model will be adjusted and the model rebuilt. As stated above the C5.0 modelling tool takes half the data to build the model and the other half to test it. It creates a new variable, the estimated value, on which to test it. We will use 'Leave' as the output, the modelling tool will create a new variable called '\$C-Leave'. If, when measuring the actual against the predicted we get an accuracy of 80% or more then we have verified the model to the level of accuracy stated.

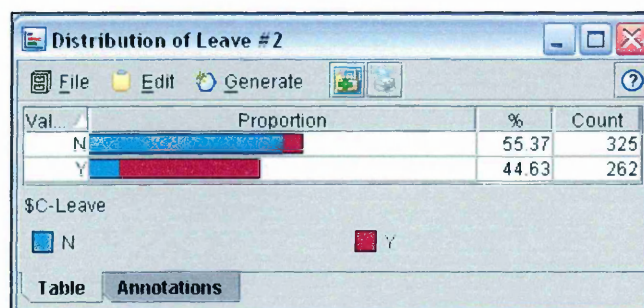
c) Build Model

Initially all variables, both demographic and problem oriented were used as inputs to the model. The variable 'Leave' was set as the model's output. This gave eighteen demographic variables and thirty-five problem oriented variables with one output variable.

The C5.0 Decision Tree modelling tool offers two types of model output, a Decision Tree Output and a Rule Based Output. An initial C5.0 model Decision Tree output model was built and run with this data. It produced a number of rules. A sample of these rules is set out below.



Clearly these rules alone are too complicated to assess and insufficient to allow decisions to be made. We will measure the created variable \$C-Leave against Leave to check the accuracy of this initial model.



The bars represent the proportion of potential 'Leavers' ('Y') and the others ('N'). The colour represents the estimated values of the variable \$C-Leave. The vast majority of these estimates appear to be correct.

The matrix below shows it more clearly.

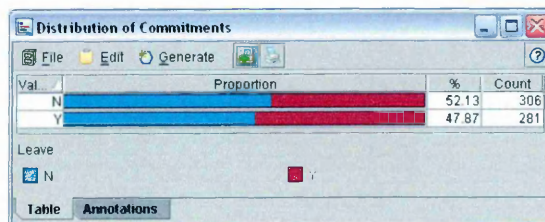
		Leave	
\$C-Leave		N	Y
N	Count	296	46
	Row %	86.550	13.450
Y	Count	29	216
	Row %	11.837	88.163

The matrix clearly shows that the percentage of potential leavers (Leave) that were correctly estimated was 88.2% (88.163) which is well above our target of 80% that we set earlier. The percentage of potential non-leavers was 86.6% (86.550) was also very high and above the 80% preset threshold. It can be concluded that we have built a relatively accurate model.

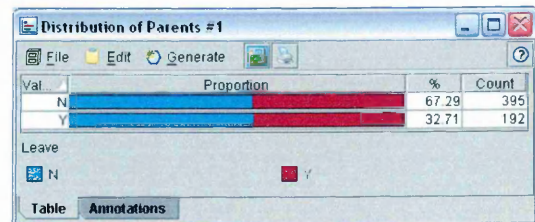
However, we now need to examine closely the variables to see if some of them (demographic and/or problem oriented) can be removed from the model as they make little or no difference to the model built. Indeed they might result in over-fitting. (See section 4.2.5 above) This process is called variable reduction. This is a vital part of the process as the smaller the number of variables to be considered the easier the exercise of assessing the 'leaving potential' will be. It has been decided to use graphical means to pursue this. Each column will be the same length and show the proportions of the different components in it.

The problem statements and demographic variables listed on the next page were removed.

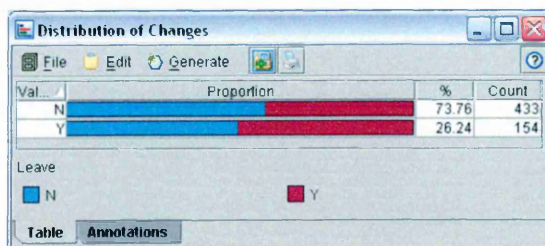
Problem Statements



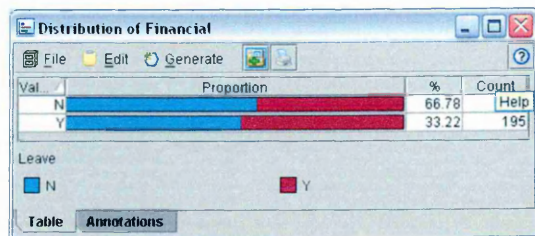
A lot of outside commitments



Lack of Parental financial support

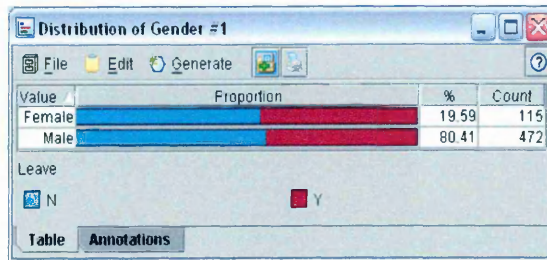


Family financial circumstances changed

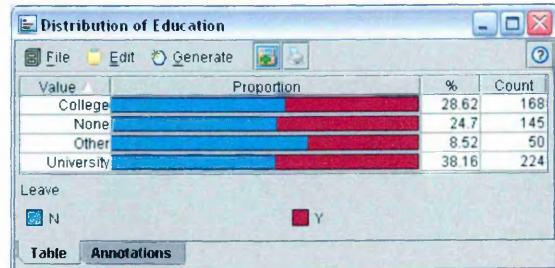


High financial commitments

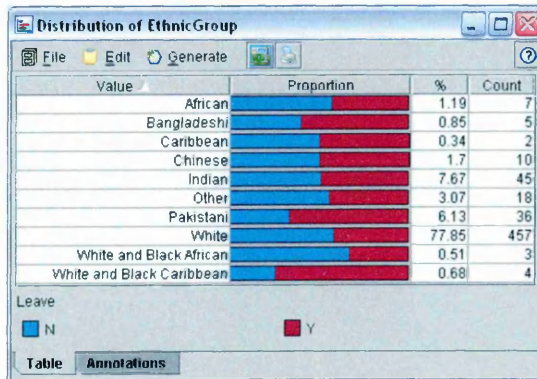
Demographic Variables



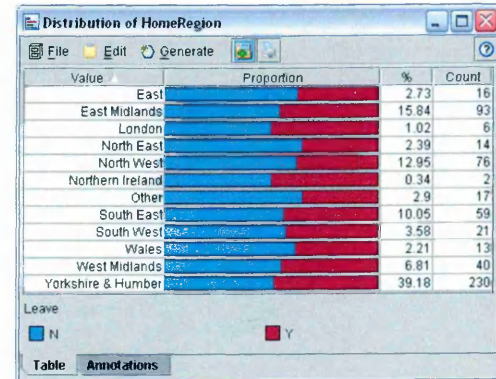
Gender



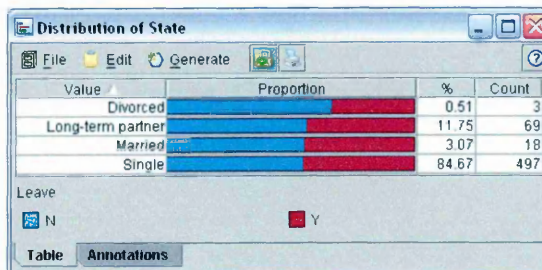
Previous Education



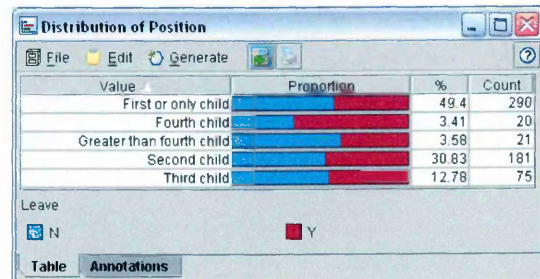
Ethnic Group



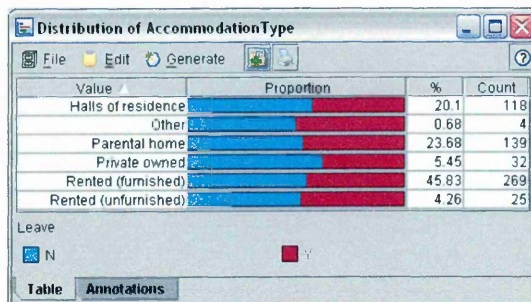
Home Region



Marital Status



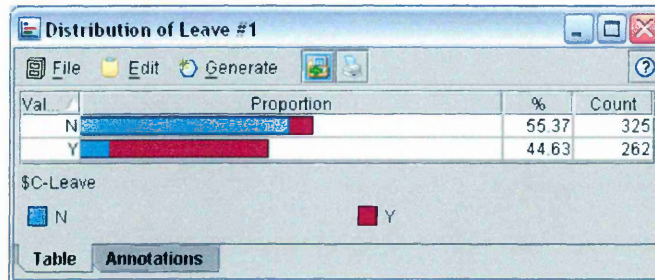
Position in Family



Type of Accommodation

The principal reason for removing the variables was because the split between 'Yes' and 'No' for the 'Leave' statement was small or because the proportion of some of the individual components were very small, for instance with Home Region. Some of the groups had a very small number of records in them. These variables were now filtered out of the input stream using a 'Filter' node in Clementine® so that they would play no further part in the data analysis.

Once this was done a new C5.0 model was built using the remaining variables as input and the 'Leave' variable (as before) as output. Looking at the output of the predicted variable '\$C-Leave' against the variable 'Leave' we get an interesting result.



The bars represent the actual variable 'Leave' and the colour the predicted variable '\$C-Leave'.

Below is the matrix for the model.

		Leave	
		N	Y
\$C-Leave N	Count	292	40
	Row %	87.952	12.048
Y	Count	33	222
	Row %	12.941	87.059

This gives an accuracy of 87.1% (87.059) which is not dissimilar to the accuracy of the original model with all the variables in. Indeed it is better at predicting potential non-leavers, 87.9% against 86.6%.

Below they are shown side by side for comparison purposes.

		Leave	
		N	Y
\$C-Leave N	Count	296	46
	Row %	86.550	13.450
Y	Count	29	216
	Row %	11.837	88.163

All variables

		Leave	
		N	Y
\$C-Leave N	Count	292	40
	Row %	87.952	12.048
Y	Count	33	222
	Row %	12.941	87.059

After variable reduction

Reducing variables is a big compromise between simplifying the model and hence making it more usable and maintaining the quality of the model. In this case we have only lost just over 1% (1.104) which is tiny. However, we have only removed a small number of variables. Seriously reducing the variables doesn't necessarily reduce the accuracy of the model because redundant variables can have an adverse effect on the outcome. They can result in over-fitting the model (see Section 4.2.5 above).

We will now build a model using only the variables that were found to be the most significant in the Market Basket Analysis exercise above. The created model is shown below, where the input variables to the Decision Tree were only Distraction, Course, Stress and Examination.

Leave			
\$C-Leave		N	Y
N	Count	253	153
	Row %	62.315	37.685
Y	Count	72	109
	Row %	39.779	60.221

Leave variables (Distraction, Course, Stress & Examination)

Clearly the accuracy has been reduced significantly to 60.2% (60.221). However it is still getting it right three out of every five times, which is better than chance!

We should continue the process of variable reduction to find an optimum solution. Variable Reduction isn't just a case of removing variables; it is removing the right variables.

d) Assess Model

We now turn to assessing the model. Earlier we chose a Decision Tree to model, but we could have used a Neural Network. We ruled it out because there were only 587 rows of data. However, it would be interesting to see if we were justified in this action. Running a Neural Network with the same inputs and output as the first Decision Tree we see the following output matrix.

Leave			
\$N-Leave		N	Y
N	Count	244	70
	Row %	77.707	22.293
Y	Count	81	192
	Row %	29.670	70.330

This shows a success rate of 70.3% (70.330) which is well below the Decision Tree model. The Neural Network matrix and the equivalent Decision Tree matrix are reproduced below, side by side for ease of comparison.

Leave			
\$N-Leave		N	Y
N	Count	244	70
	Row %	77.707	22.293
Y	Count	81	192
	Row %	29.670	70.330

Neural Network (All variables)

Leave			
\$C-Leave		N	Y
N	Count	296	46
	Row %	86.550	13.450
Y	Count	29	216
	Row %	11.837	88.163

Decision Tree (All variables)

Since the 70.3% obtained from the Neural Network is well below the 80% we set ourselves, and also taking into account that Decision Tree gave us 88.2% then we were indeed justified in our action of rejecting a Neural Network model. However it is far from unsatisfactory. It does predict the correct potential leavers with over 70% success.

We now move back to assessing the Decision Tree models as a whole. We preset our criterion with a success rate of 80% in 9.4.2(b). If we examine the three different models that were built we see that two have passed out stated criterion and one has failed. The first model uses ALL the variables (except 'Leave') as input. After variable reduction the second model is somewhat simpler without compromising the accuracy significantly.

However, the Decision Tree model that was built after considering the major relationships from the Rule Association modelling tool (GRI) must be considered is a 'bridge too far'. Though it is simple, with only four predictors (Distraction, Course, Stress & Examination) it gives an unacceptable level of accuracy of 60.2% against the original 88.2%. Our optimum solution is going to lie somewhere between these extremes of variable reduction.

The first iteration of the variable reduction is to take all nine of the primary problems from the GRI model and use them as inputs to the model still maintaining 'Leave' as the output, to see if the model improves.

		Leave	
\$C-Leave		N	Y
N	Count	255	102
	Row %	71.429	28.571
Y	Count	70	160
	Row %	30.435	69.565

C5.0 using the nine primary problems

The problems were Commitments, Confidence, Course, Different, Distractions, Examinations, Management, Stress and Work.

An accuracy of 69.6% (69.565) is rather better than the earlier attempt of 60.2% but still needs some work. The next step is to introduce some of the demographic variables. The ones that were initially excluded in the variable reduction were Gender, Education, Ethnic Group, Home region, State, Position and Accommodation Type. So we will introduce the others, Induction, Socio Economic Group, Age Group, Sibling, Background, Work Experience, Relevant, Year Description.

		Leave	
\$C-Leave		N	Y
N	Count	288	104
	Row %	73.469	26.531
Y	Count	37	158
	Row %	18.974	81.026

C5.0 using nine primary problems and eight primary demographics

This model has an accuracy of 81.0% (81.026) which is within our acceptable limit of 80%. Further variable reduction is pointless as it will almost certainly reduce the accuracy to below the 80% level.

It is therefore concluded that the optimum decision tree (C5.0) model has been found with eight demographic input variables and nine input problem variables. This is much simpler than the original model of eighteen demographic variables and thirty-six problem variables. The accuracy of the model has only been reduced from 88.2% to 81.0% which still remains within our preset acceptance.

Examining the Demographic Profile

All that now remains is to profile the students at most risk of leaving. So taking each of the remaining demographic categories we examine the relative proportions of each of the specific groups as shown in the matrices below.

		Leave	
AgeGroup		N	Y
21 to 25	Count	187	143
	Row %	56.667	43.333
26 to 30	Count	14	13
	Row %	51.852	48.148
31 to 40	Count	14	5
	Row %	73.684	26.316
Over 40	Count	4	2
	Row %	66.667	33.333
Under 21	Count	106	99
	Row %	51.707	48.293

Potential leavers by Age Group

		Leave	
SocioEconomicGroup		N	Y
Managerial	Count	66	56
	Row %	54.098	45.902
Partly skilled	Count	11	18
	Row %	37.931	62.069
Professional	Count	134	95
	Row %	58.515	41.485
Skilled manual	Count	62	39
	Row %	61.386	38.614
Skilled non-manual	Count	15	12
	Row %	55.556	44.444
Technical	Count	23	19
	Row %	54.762	45.238
Unskilled	Count	14	23
	Row %	37.838	62.162

Potential leavers by Socio Economic Group

		Leave	
Education		N	Y
College	Count	95	73
	Row %	56.548	43.452
None	Count	78	67
	Row %	53.793	46.207
Other	Count	32	18
	Row %	64.000	36.000
University	Count	120	104
	Row %	53.571	46.429

Post-school Education of Parents

		Leave	
Background		N	Y
College	Count	215	174
	Row %	55.270	44.730
Other	Count	2	4
	Row %	33.333	66.667
Polytechnic	Count	3	3
	Row %	50.000	50.000
School only	Count	79	66
	Row %	54.483	45.517
University	Count	26	15
	Row %	63.415	36.585

Potential leavers by Previous Education

Leave			
Sibling		N	Y
More than three	Count	35	35
	Row %	50.000	50.000
None	Count	40	24
	Row %	62.500	37.500
One	Count	140	104
	Row %	57.377	42.623
Three	Count	29	35
	Row %	45.312	54.688
Two	Count	81	64
	Row %	55.862	44.138

Potential leavers by Number of Siblings

Leave			
WorkExperience		N	Y
1 to 2 years full-time	Count	88	71
	Row %	55.346	44.654
More than 2 years full-time	Count	45	22
	Row %	67.164	32.836
None	Count	22	17
	Row %	56.410	43.590
Part-time only	Count	124	120
	Row %	50.820	49.180
Under 1 year full-time	Count	48	32
	Row %	59.974	41.026

Potential leavers by Work Experience

Leave			
YearDescription		N	Y
First	Count	61	54
	Row %	53.043	46.957
Fourth	Count	133	82
	Row %	61.860	38.140
Over fourth	Count	7	5
	Row %	58.333	41.667
Second	Count	58	69
	Row %	45.669	54.331
Third	Count	66	52
	Row %	55.932	44.068

Potential leavers by Year Group

Leave			
Relevant		N	Y
No	Count	124	134
	Row %	48.062	51.938
Unsure	Count	21	16
	Row %	56.757	43.243
Yes	Count	180	112
	Row %	61.644	38.356

Potential leavers by Relevant Job Experience

Age Group

The most vulnerable age groups are the Under 21's and 26-30's with over 48% of the students potentially leavers. This is not surprising since these are the bulk of the students. This information should be looked at with caution as some of the groupings were very small. For instance, although the 26-30 age group also gave a figure of over 48% there were only 27 students in the sample. In retrospect the author believes that the students should have been asked their actual age rather than their age group.

Socio Economic Group

By far the most vulnerable groups are those from unskilled and partly skilled backgrounds, both with 62% of potential leavers. None of the other groups come anywhere near them.

Post-school Education of Parents

46% of students from parents with no post-school education considered leaving. However, the same percentage of students from parents with a university education also considered leaving. This is an interesting point but beyond the scope of this study.

Number of Brothers and Sisters

It appears that the students from larger families are more likely to consider leaving, those students with no brothers and/or sisters being significantly less likely (37.5%).

Previous Educational Experience

Those from 'school-only' were the most likely to consider leaving with 45.5%.

Work Experience

Those with more than two years work experience were significantly less likely to consider leaving (32%) than any other group. All the rest were over 40%.

Relevant Work Experience

Those with no relevant work experience were significantly more likely to consider leaving (52%).

Year in University

The most vulnerable year group was the second year with more than 54% that had considered leaving. This was followed by the first year with 47%. The higher the year the less vulnerable the students appeared to be.

Examining the Problem Issues

It was shown, when examining Rule 1 (section 9.4.1) that the students that were most at risk had expressed problems in the following areas:

- Different
- Examinations
- Commitments
- Stress
- Confidence
- Distractions
- Management
- Course
- Induction

For completeness these are set out in the table below as matrices:

Leave			
Different		N	Y
N	Count	187	85
	Row %	68.750	31.250
Y	Count	138	177
	Row %	43.810	56.190

3. Course different from expectations

Leave			
Examinations		N	Y
N	Count	171	108
	Row %	61.290	38.710
Y	Count	154	154
	Row %	50.000	50.000

11. Examination performance below expectations

Leave			
Commitments		N	Y
N	Count	176	130
	Row %	57.516	42.484
Y	Count	149	132
	Row %	53.025	46.975

17. Outside commitments high

Leave			
Stress		N	Y
N	Count	157	99
	Row %	61.328	38.672
Y	Count	168	163
	Row %	50.755	49.245

20. Stress problems

Leave			
Confidence		N	Y
N	Count	239	158
	Row %	60.202	39.798
Y	Count	86	104
	Row %	45.263	54.737

22. Lack of self confidence

Leave			
Management		N	Y
N	Count	239	172
	Row %	58.151	41.849
Y	Count	86	90
	Row %	48.864	51.136

30. Bad at managing finances

Leave			
Induction		N	Y
N	Count	252	160
	Row %	61.165	38.835
Y	Count	73	102
	Row %	41.714	58.286

36. The induction didn't help to feel more comfortable

Leave			
Distractions		N	Y
N	Count	211	131
	Row %	61.696	38.304
Y	Count	114	131
	Row %	46.531	53.469

24. Too many distractions that affect ability to study

Leave			
Course		N	Y
N	Count	140	65
	Row %	68.293	31.707
Y	Count	185	197
	Row %	48.429	51.571

32. Have sometimes found course stressful

It can be seen that the lowest percentage of students that had a 'Yes' for **Leave** problems and a 'Yes' for the specified one is 47% (46.975) with **Commitments**. Apart from **Stress**, all the rest have 'Yes' with 'Yes' in a majority, suggesting a strong connection. The 'No' with 'No' is even stronger with all but two of them being over 60% correlation.

Summary

The tables below set out in brief the above findings:

Demographic Issue	Vulnerable Group
Age-group	Under 21
Socio-economic Group	Unskilled or Partly Skilled
Parents post-school education	None or University
Number of brothers & sisters	Three or more
Previous educational experience	School only
Work experience	Less than 2 years
Relevant work experience	None
Year in University	Second, First

Issue	Description
Different	Course different from expectation
Examinations	Examination performance below expectations
Commitments	Outside commitments high
Stress	Stress problems
Confidence	Lack self confidence
Distractions	Too many distractions that affect ability to study
Management	Bad at managing finances
Course	Have sometimes found course stressful
Induction	The induction didn't help to feel more comfortable at University

These being associated with the output variable:

Issue	Description
Change	Considered changing/leaving at some stage

9.4.3 Rule 3

Determine clusters of students with similar response patterns and ascertain which cluster(s) are most at risk of attrition

a) *Select Modelling Technique*

The wording of this goal suggests that a clustering model such as Kohonen Network or K-Mean is suitable. As suggested in Section 9.1(d) above, it was decided to use Kohonen Networks as there was little to choose between them and the author had more experience of Kohonen Networks.

b) *Generate Test Design*

Again we will use the in-built mechanism for testing as we did with the previous two models. The modelling tool was run with default values and various numbers of clusters. For display purposes the model builds the clusters in the form of a matrix, using rows and columns as 2-dimensional coordinates of the cluster cells.

c) *Build Model*

The clusters are defined by rows and columns as shown below:

(0,2)	(1,2)	(2,2)	(3,2)
(0,1)	(1,1)	(2,1)	(3,1)
(0,0)	(1,0)	(2,0)	(3,0)

This is in accordance with the normal Mathematics co-ordinates. The first number represents the column and the second represents the row. A 4 by 3 model was first created and the sizes of the clusters were examined as follows:



Building a Kohonen Network with 12 clusters (4x3)

This gave the following output graph:

Val...	Proportion	%	Count
00		12.78	75
01		5.62	33
02		10.9	64
10		7.33	43
11		3.07	18
12		6.98	41
20		6.81	40
21		3.58	21
22		6.98	41
30		10.73	63
31		6.47	38
32		18.74	110

Data points distribution from a 4x3 Cluster combination

Examining this found that cluster (0,0), (0,2), (3,0) and (3,2) appeared to have the highest number of records in them. However, the other clusters were far from insignificant. Since only 587 rows of data were available it was felt that increasing the number of clusters would only worsen the situation. A number of other models were built with different cluster combinations and finally the 3 by 3 matrix was chosen as the most stable. This is documented below.

(0,2)	(1,2)	(2,2)
(0,1)	(1,1)	(2,1)
(0,0)	(1,0)	(2,0)



Building a Kohonen Network with 9 clusters (3x3)

The clusters were then plotted on a graph as before and gave the following distribution:





Val...	Proportion	%	Count
00		19.93	117
01		7.5	44
02		14.48	85
10		9.03	53
11		2.73	16
12		6.64	39
20		15.16	89
21		7.67	45
22		16.87	99

This appears to be rather more definite. There are four clear clusters that are much larger than the rest. These are (0,0), (0,2), (2,0) and (2,2). These

were isolated out by using a **Select** node. This removes the other clusters so that you can concentrate on the chosen ones. The formula used was:

cluster = "00" or cluster = "22" or cluster = "02" or cluster = "20"

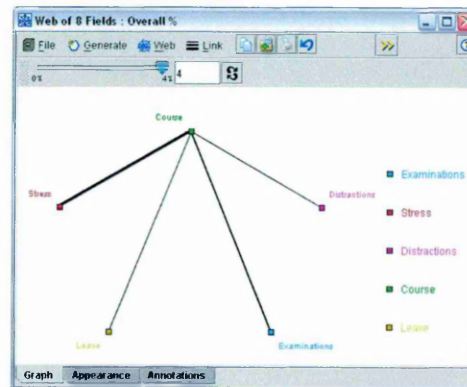
This is shown in the graph below:

Val...	Proportion	%	Count
00		30.0	117
02		21.79	85
20		22.82	89
22		25.38	99

The clusters now need to be analysed to find their demographic and problem makeup.

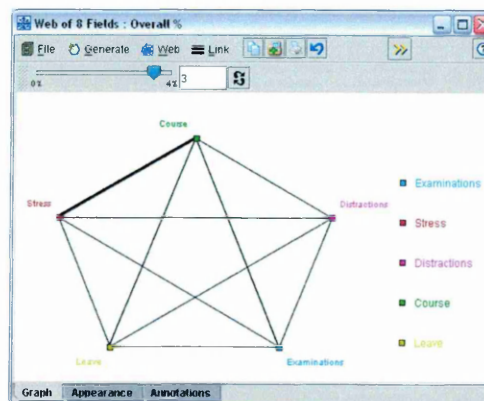
d) Assess Model

We will start by looking at the interrelationships between the clusters and the strongest variables. A cluster diagram is the best starting point. We will first look at the strongest relationships:



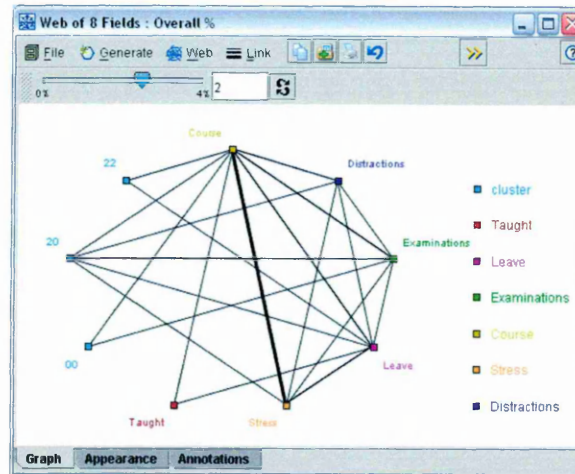
Web showing the strongest relationships

The variables Stress, Course, Distractions, Examinations and Leave are the same major problems that were identified by the Apriori model in Section 9.4.1(c) above. If we move the blue slide on the top left across slowly we will gradually see more relationships appear. The thicker the line is, the stronger the relationship. **Stress** and **Course** are clearly the strongest relationship.



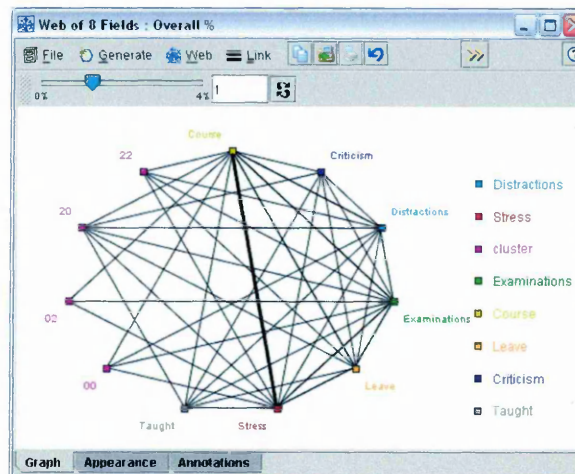
Web showing the same variables above, but now fully connected

The above web diagram shows the same variables, but now they are all connected to each other. There seems to be little difference in the thickness of the lines though the new ones are a little less strong than the ones shown on the previous diagram.



Web showing the three strongest clusters and additional problem variables

Three of the clusters have now appeared. Cluster (0,0) is connected to Course and Examinations. Cluster (2,0) is connected to Stress, Leave, Examinations Distractions and Course. Cluster (2,2) is connected to Course and Leave. This information will form the basis of the analysis that follows.



Web showing all the clusters and the major problem variables

This web clearly shows the major effects of the principal problem variables on the different cluster groups. The final web need not be shown as it connects all the above variables together without introducing anymore problems. More use will be made of the above relationships in the later analysis.

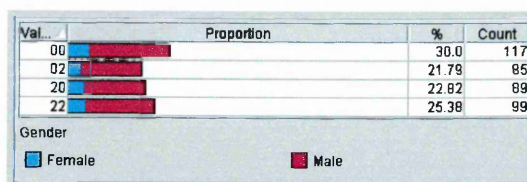
We shall now look at the demographic makeup of the clusters. Below is documented only the demographics that showed a clear distribution pattern.

Gender

The matrix and graph below show the gender split across the clusters. Although cluster (0,2) contains fewer females, it isn't significantly different.

Gender		
cluster	Female	Male
00	20.513	79.487
02	17.647	82.353
20	21.348	78.652
22	20.202	79.798

Matrix of Gender split by Cluster



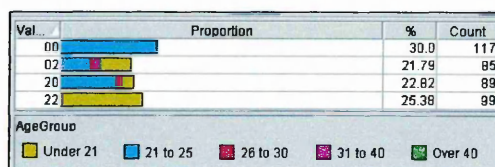
Graph of Gender split by Cluster

Age Group

The next matrix shows the age-group split:

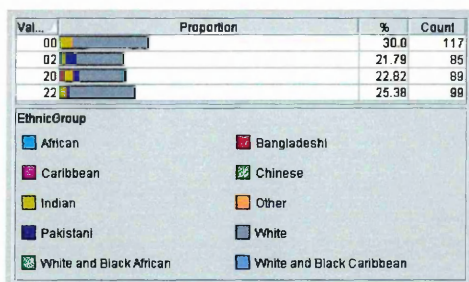
AgeGroup					
cluster	21 to 25	26 to 30	31 to 40	Over 40	Under 21
00	100.000	0.000	0.000	0.000	0.000
02	41.176	4.706	9.412	3.529	41.176
20	75.281	7.865	2.247	1.124	13.483
22	2.020	1.010	0.000	0.000	96.970

This clearly shows that cluster (0,0) is exclusively 21-25 year olds and cluster (2,2) is almost exclusively the Under 21s. (2,0) is mixed, but primarily 21-25 year olds with a smattering of other groups. The older students are almost exclusively located in clusters (0,2) and (2,0), cluster (0,2) being the most mixed. This can be seen more clearly by looking at a graph below.

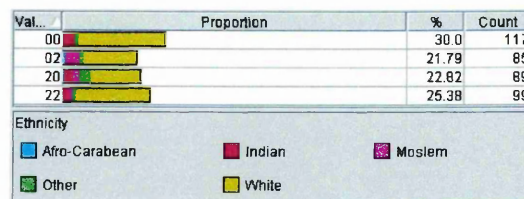


Ethnic Groups

Looking at the graph below it can be seen that most of the Ethnic groupings are very small. However some patterns are emerging that might be worthy of further investigation. The left-hand graph shows the separate groupings and the right-hand graph shows a reduced number of categories. Pakistani and Bangladeshi students (Moslem) have been grouped together as have African and Caribbean (Afro-Caribbean).



Cluster split by Ethnic Group

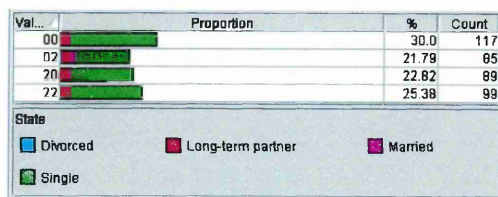


Cluster Split by Ethnic Group (reduced)

The reduced category graph (on the right) clearly shows that most of the Moslem students are located in cluster (0,2) and Indian students in cluster (0,0). The largest grouping of white students is in cluster (0,0) with almost as many on cluster (2,2). However there are still large numbers in the other clusters. Most of the others ethnic groups are located in cluster (2,0).

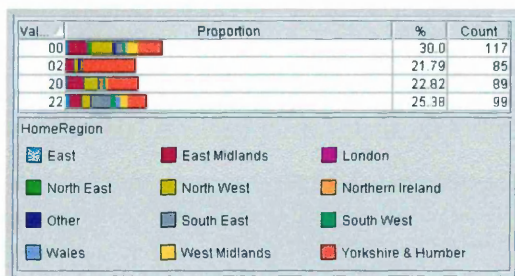
Marital Status

Apart from most of the married students being located in cluster (0,2), there appears to be little else of interest in this.

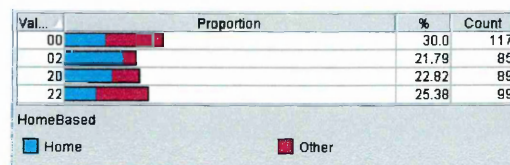


Home Region

The right-hand graph regroups the students into locally-based students (Yorkshire & Humber, East Midlands and North West). This clearly shows that cluster (0,2) is largely students that are locally based, there being very few non-locally based students in this cluster.

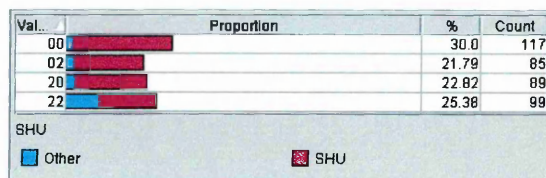


Cluster by Home Region



Cluster by Home Region (reduced)

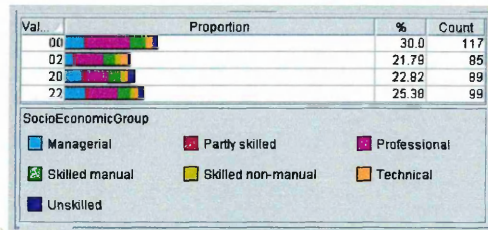
Because 15% of the students in the survey were from other universities it is difficult to derive much else from this. Below is a graph of the SHU and non-SHU against the clusters to see their distribution.



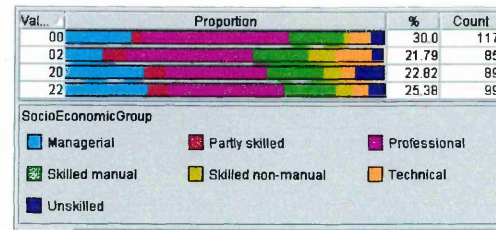
This shows that most non-SHU students are located in cluster (2,2), suggesting that there are some differences between SHU students and others surveyed. The analysis of this is beyond the scope of this survey but might be worth investigation in an additional study.

Socio-Economic Group

On first sight (left-hand graph), the different grouping in each cluster appear to be similar. However, on closer examination there seems to be a lower preponderance of students from managerial background in cluster (0,2) and more students from Professional backgrounds in cluster (0,0). The right-hand graph shows the same data, but the columns have been drawn of equal length for ease of comparison. Examining this shows a higher proportion of students from Technical backgrounds in cluster (0,2).



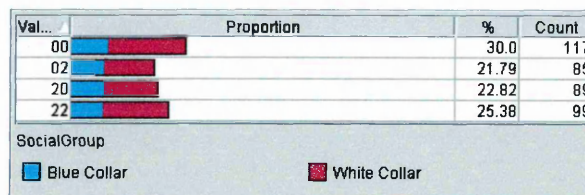
Cluster by Home Region



Cluster by Home Region (reduced)

SocioEconomicGroup							
cluster	Managerial	Partly skilled	Professional	Skilled man...	Skilled non-...	Technical	Unskilled
00	20.513	3.419	46.154	17.094	1.709	6.838	4.274
02	11.765	7.059	40.000	17.647	8.235	11.765	3.529
20	24.719	6.742	31.461	17.978	5.618	4.494	8.989
22	26.263	6.061	36.364	16.162	5.051	5.051	5.051

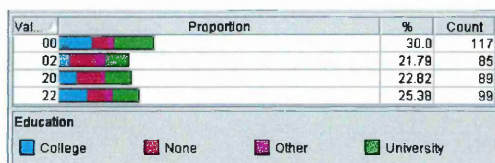
In the graph below the grouping were reduced to 'Blue Collar' and 'White Collar'. This shows no real conclusive evidence of anything except perhaps (0,0) has more White Collar backgrounds.



White Collar was defined as Professional and Managerial and Blue Collar the rest.

Education

This category was the educational background of the main wage earner of the family. It shows that those with no post-school education were more strongly associated with cluster (0,2) whereas those with a college or university education were more strongly related to cluster (0,0).



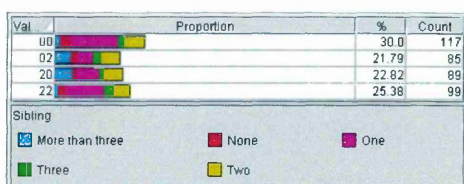
Cluster by Parental Education

Education				
cluster	College	None	Other	University
00	35.043	17.094	5.983	41.880
02	14.118	37.647	14.118	34.118
20	23.596	32.584	6.742	37.079
22	35.354	22.222	10.101	32.323

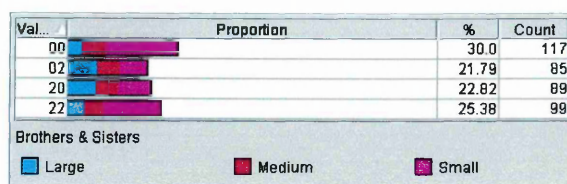
Cluster by Parental Education

Siblings

It is clear from the left-hand graph below that those students from smaller families (one or no brothers or sisters) are principally located in clusters (0,0) and (2,2). Further analysis is hampered because of the number of categories. If these are reduced to Small (one or less), Medium (two) and Large (greater than two) we get the right-hand graph showing that clusters (0,2) and (2,0) contain more of the larger families than the other two. Clusters (0,0) and (2,2) have a majority of small families in them.



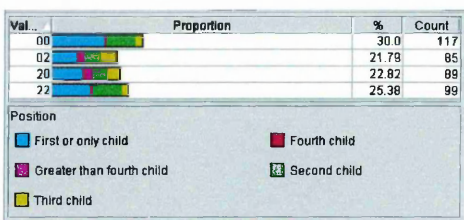
Cluster by Number of Siblings



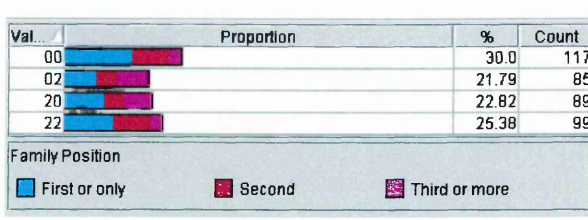
Cluster by Family Size

Position in Family

The position in the family is a bit less clear. As stated earlier in 9.2(e), there appears to be a discrepancy between number of siblings and position in family which is likely to be accounted for if step brothers and sisters are taken into consideration in number of siblings and not in position in family.



Cluster by Position in Family

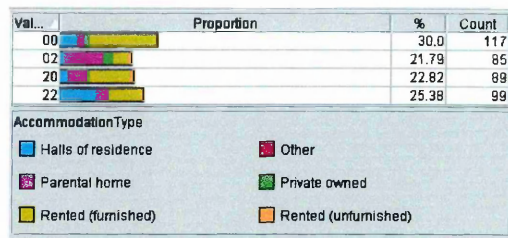


Cluster by Position in Family (regrouped)

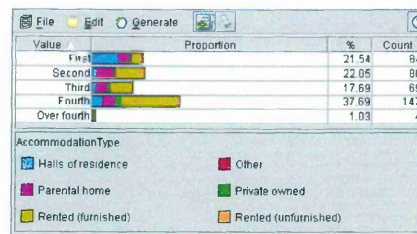
The right-hand graph shows a distribution with three rather than five groups. It shows that although all cluster contain all positions, cluster (0,0) has a higher number of 'First or only' children, the cluster (0,2) many of the 'Third or More', though these are almost as many in cluster (2,0). Cluster (2,2) contains more 'Second' child though outnumbered by 'First or only child'. However, on balance there is no overwhelming certainly.

Type of Accommodation

Looking at the left-hand graph below, the Type of Accommodation is much more definite. Cluster (0,0) contains mainly students from rented furnished accommodation. These are generally single students above first year. First year students tend to live in Halls of residence. To explore this further another graph is needed that shows the demographic split by year of study.



Cluster by Accommodation Type

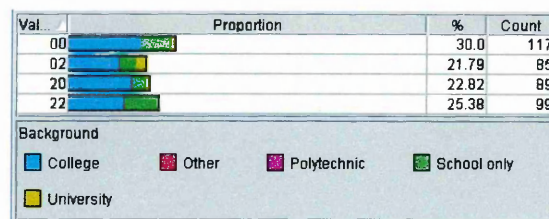


Accommodation Type by Year of Student

The right-hand graph shows student year by accommodation type. This clearly shows that First year students favour Halls of Residence whereas students from second year upwards prefer Rented Furnished accommodation. Students living in parental home seem fairly equally distributed throughout the year groups though for some reason second year students seemed rather more likely to. It also appears that some fourth year students (mainly those returning from work placement) prefer to move back into Halls of Residence, though this is a small proportion of the group.

Previous Educational Background

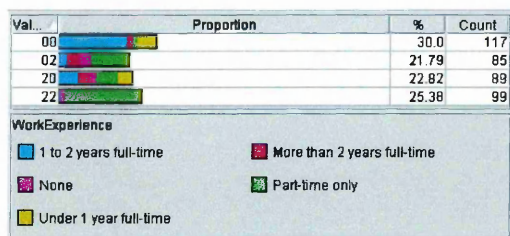
Most students surveyed had come from a college background. Those students that had come directly from school were mainly located in cluster (2,2), though a fair number were in cluster (0,0). Those from a college background were slightly more likely to be in cluster (0,0). It has to be said though, that there is no really conclusive demographic split.



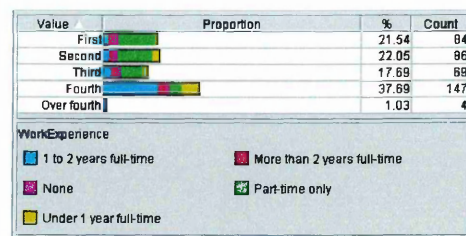
Cluster by Previous Educational Background

Work Experience

When looking at student work experience in the left-hand graph below, it was found that cluster (2,2) was mainly made up of students who had only part-time work experience. On the other hand cluster (0,0) was made up of mainly students with 1 to 2 years full-time experience. Those with under a year seem to be split between clusters (0,0) and (2,0).



Cluster by Work Experience

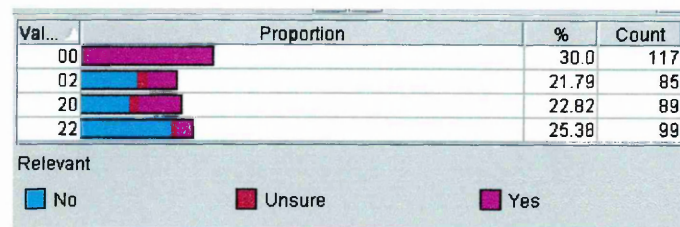


Year Group by Work Experience

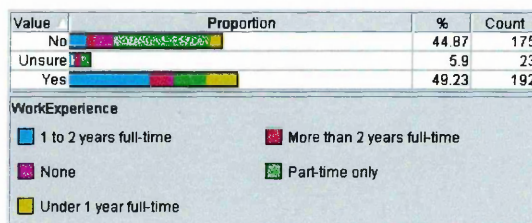
The right-hand graph shows the distribution of work experience across the different year groups. This shows very clearly the majority of fourth year students had one to two years of work experience. This reflects the fact that many of these students had undergone a one year placement in their third year. However it does show that the vast majority of students had worked at least part-time.

Relevant Work Experience

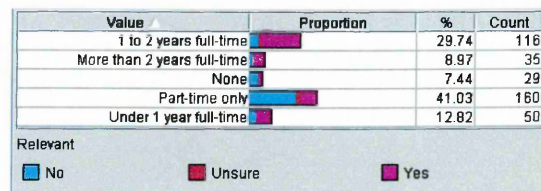
When students were asked about whether they had work experience relevant to their course, cluster (0,0) was overwhelmingly dominated by those students that had, whereas cluster (2,2) was largely students with no relevant work experience.



Reviewing work experience against relevant work experience gave a very interesting picture. The two graphs below show the two issues plotted, first with relevant experience against amount and secondly the other way round.



Relevant experience against type of experience



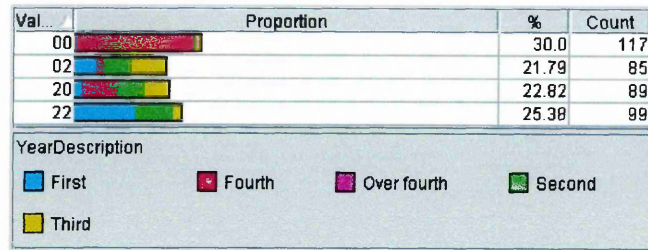
Type of experience against relevant experience

Looking at these two graphs it shows a clear pattern that those with no experience relevant to the course had mainly part-time work experience whereas the majority of those with experience relevant to the course had at least one year experience. Viewed the other way round, part-time work tended to be irrelevant to the course whereas full-time work tended to be relevant.

Post Script: A small number of those students that had said they had no work experience said that it (this non-experience) was relevant to their course. This, though minor, does show the importance of using cross-validation questions and making allowances for inaccuracies in the data.

Year

Examining clusters by year of study shows a relatively clear picture. Cluster (0,0) was almost exclusively fourth year students whereas cluster (2,2) was predominantly first year students. Third year students were mainly located in cluster (0,2) with a fair number in (2,0). Second year students were split between (0,2), (2,0) and (2,2).



Summary

The tables below set out in brief the above findings:

	Cluster (0,0)	Cluster (0,2)	Cluster (2,0)	Cluster (2,2)
Age-group	21 to 25 years			Under 21 years
Home Region		Local to university		
Number of Brothers & Sisters	One sibling or less			
Position in family	First or only child			
Accommodation Type	Furnished	Parents home	Furnished (mainly)	Halls (mainly)
Work Experience	1-2 years Full-time	< 1 year full-time	< 1 year full-time	Part-time only
Relevant work experience	Relevant experience			No relevant experience
Year of study	4 th year	2 nd & 3 rd year	Mainly 2 nd & 3 rd year	1 st year
Stress problems	Yes	Yes	Yes	Yes
Too many distractions that affect ability to study	Yes		Yes	Yes
Examination performance below expectations	Yes	Yes	Yes	Yes
Have sometimes found the course stressful	Yes	Yes	Yes	Yes
Criticism disliked			Yes	
Taught badly			Yes	
Considered changing or leaving at some stage			Yes	Yes

9.5 Evaluation

9.5.1 Introduction

Having now built three models and analysed each of them they now need to be compared against each other to find commonality and differences. Any anomalies will then need to be investigated to ensure that no errors have been made. Any remaining differences then need to be accounted for wherever possible.

9.5.2 What has been found

What follows has been lifted from the summary of each of the rule summaries above.

Rule 1: Determine interrelationships within the response data

This rule was investigated by Rule Association (Apriori) and the variable relationships below were found to be the strongest at a confidence level of 70% and support of 25%:

Rate	Problem
High	Have sometimes found course stressful (Course)
High	Stress problems (Stress)
	Course different from expectation (Different)
	Lack self confidence (Confidence)
	Outside commitments high (Commitments)
High	Too many distractions that affect ability to study (Distractions)
	Behind with work (Work)
High	Examination performance below expectations (Examinations)
	Personal circumstances changed (Circumstances)
	Bad at managing finances (Management)
High	Considered changing/leaving at some stage (Leave)

Rule 2: Predict the likelihood that a student will leave given other known factors

The tables below set out in brief the above findings:

Demographic Issue	Vulnerable Group
Age-group	25 & under
Socio-economic Group	Unskilled or Partly Skilled
Parents post-school education	None or University
Number of brothers & sisters	Three or more
Previous educational experience	School only
Work experience	Less than 2 years
Relevant work experience	None
Year in University	Second, First

Issue	Description
Different	Course different from expectation
Examinations	Examination performance below expectations
Commitments	Outside commitments high
Stress	Stress problems
Confidence	Lack self confidence
Distractions	Too many distractions that affect ability to study
Management	Bad at managing finances
Course	Have sometimes found course stressful
Induction	The induction didn't help to feel more comfortable at University

These being associated with the output variable:

Issue	Description
Change	Considered changing/leaving at some stage

Rule 3: Determine clusters of students with similar response patterns and ascertain which cluster(s) are most at risk of attrition

The tables below set out in brief the above findings:

	Cluster (0,0)	Cluster (0,2)	Cluster (2,0)	Cluster (2,2)
Age-group	21 to 25 years			Under 21 years
Home Region		Local to university		
Number of Brothers & Sisters	One sibling or less			
Position in family	First or only child			
Accommodation Type	Furnished	Parents home	Furnished (mainly)	Halls (mainly)
Work Experience	1-2 years Full-time			Part-time only
Relevant work experience	Relevant experience			No relevant experience
Year of study	4 th year	2 nd & 3 rd year	Mainly 2 nd & 3 rd year	1 st year
Stress problems	Yes	Yes	Yes	Yes
Too many distractions that affect ability to study	Yes		Yes	Yes
Examination performance below expectations	Yes	Yes	Yes	Yes
Have sometimes found the course stressful	Yes	Yes	Yes	Yes
Criticism disliked			Yes	
Taught badly			Yes	
Considered changing or leaving at some stage			Yes	Yes

9.5.3 Summary

All the rules were investigated by a different Data Mining model as follows:

Rule 1: Rule Association (Apriori)

Rule 2: Decision Tree (C5.0)

Rule 3: Clustering (Kohonen Network)

There was found to be a great deal of commonality between the different analyses. Set out in the table below are the student problems that were found by each of the models. The problems are set out in rows and the models in columns.

Problem	Rule 1	Rule 2	Rule 3
Have sometimes found course stressful	Yes	Yes	Yes
Stress problems	Yes	Yes	Yes
Too many distractions that affect ability to study	Yes	Yes	Yes
Examination performance below expectations	Yes	Yes	Yes
Considered changing/leaving at some stage	Yes	Yes	Yes
Bad at managing finances	Yes	Yes	
Course different from expectation	Yes	Yes	
Lack self confidence	Yes	Yes	
Outside commitments high	Yes	Yes	
Behind with work	Yes		
Personal circumstances changed	Yes		
The induction didn't help to feel more comfortable at University		Yes	
Criticism disliked			Yes
Taught badly			Yes

In all there are five student problems that were found by all the models. These were Course, Stress, Distraction, Examinations and Leave. There was even more conformity between the analysis for Rule 1 and Rule 2.

Turning now to the demographic data, the relationships were not found to be as strong. However when related to the major clusters significant patterns have emerged. The major demographic features that appear to affect student attrition are as follows:

Demographic	Vulnerable group	Less-vulnerable
Age Group	Under 21	21 and over
Year of Study	2 nd Year, 1 st Year	4 th Year, 3 rd Year
Home Region	Non-local	Local students
Family size	Large (3 or more)	One sibling or less
Position in family	Younger sibling	First or only child
Accommodation	Halls of residence	Furnished rented or parents home
Work Experience	Part-time only or none	At least 1 year full-time
Relevant Experience	No relevant work experience	Relevant work experience

9.5.4 Vulnerable Group

Drawing all the above findings together it is concluded that the most vulnerable group (the most likely to leave) have the following characteristics:

Antecedents

Strongest (from at least two rules)

Under 21
2nd Year, 1st Year
Non-local
From a large family (3 or more children)
Younger sibling
Halls of residence
Have no work experience or only part-time
Have no relevant work experience
Have sometimes found the course stressful
Have had stress problems
Too many distractions that affect ability to work
Examinations performance below expectations
Bad at managing finances
Course different from expectation
Lack self confidence
Outside commitments high

Weaker (from only one rule)

Behind with work
Personal circumstances changed
Induction didn't help to feel more comfortable
Criticism disliked
Taught badly

Consequence

Considered leaving/changing at some stage

All the factors measured different issues, though there might have been some cross-over between them, for instance 'general stress' and 'course stress' as a result of difficulties with the course.

9.6 Deployment

These results will be used to formulate a template for a set of semi-structured interviews to be conducted with a small cross-section of the SHU research population. This is designed to answer Objective 4, *"Validate the results of the quantitative research within ITPA by means of structured interviews"*. This is accounted for in Appendix K.

Success Criterion

It is useful to remember here the success criterion that was set in section 9.1(a).

'Give useful insights into the relationships between the student demographics and the associated student problems encountered'.

Chapter 10: Focus Group Recommendations

10.1 Introduction

The Faculty of Arts, Computing, Engineering and Sciences (ACES) has a Retention Group (FARG) of which the author is a member. A Focus Group of six members drawn largely from this group was convened to discuss the findings of the study and compile a set of recommendations designed to be put to the Faculty.

At the first meeting of the Focus Group a background to the study was given together with the method and form of data recorded. The chosen data analysis method (Data Mining) was explained and justified and a list of findings issued.

10.2 Discussion of Findings

The list of findings was perused and digested before discussion of actions was undertaken. The discussion opened with a debate about whether students should be encouraged to adapt to the ethos of university or whether the university should adapt to the changing profile of the students. It was generally thought that much of university practice was still geared towards the 'Middle Class' ideals which excluded many students coming from other backgrounds. It was thought that this might be working against moves towards widening participation.

The group discussed the current faculty proposal for an extended induction for first year students and what additional recommendations might be put in place. Since the students most at risk were found to be second and first years this seemed an appropriate vehicle for some of the issues.

Lack of work experience, and in particular relevant work experience was discussed first. It was thought that the students should be identified early, perhaps during the induction period and pointed to the Careers Services to look at possible relevant part-time work in the city or within the university. Apart from giving valuable relevant work experience it would also help to alleviate financial problems.

The Student Halls of Residence were then discussed. The students, who lived in these, whatever their year group, were found to be more at risk. It was thought that this was a University Issue rather than one to be dealt with within

the faculty, as the students in halls were from the full spectrum of faculties. However, it was felt that the Students Union would be a good vehicle to use here. The encouragement of Halls activities such as Rag Week and reviews were discussed.

Since a high percentage of SHU students were home-based and were found to be less at risk than non-local students it was suggested that recruitment should focus on local students. Further research might be needed to look at the connection between local students and living in the parental home.

A discussion ensued concerning the younger siblings of large families. It was appreciated that it was not the purpose of the study to find the reasons behind these being a vulnerable group but further research in this area might be fruitful. It was thought that the counselling service might be approached for guidance in this area.

The subject of Student Stress was discussed. It was pointed out that stress was often attributed to factors outside the university as well as inside. However, stress related to the course could be a by-product of general stress and vice versa. Involving the University Counselling service, say in induction might be a great advantage.

Since both first and second year students were in the at-risk group it was thought that second year mentoring of first year students might be worth looking at. This might give a stronger sense of belonging and integration for both groups of students. It was also thought that non-first years residing in Halls of Residence might be used profitably as 'Senior Students' or mentors. Another suggestion was to look at more vertical structures.

Examinations were discussed openly. It was generally thought that the examination system needed to be looked at, in particular a study into a comparison of results from coursework and examinations to see if there was a strong enough correlation between them. Anecdotal evidence suggested that examination results were generally below those of coursework. The style of questions was discussed, in particular the content within questions, using different parts of questions to ramp up the degree of difficulty.

Too many outside interests were seen to be a strong de-motivating influence on students, often resulting in student lack of progress. However, an instance of a

Chapter 10
student giving up all outside interests when coming to university was cited. The student believed that this was a major factor in him failing his first year.

The link between lack of self-confidence, socio-economic group and stress was thought to be worth an investigation. The author said that he might be able to look at this in the current dataset, but that another study might be needed. A look at socio-economic groups and home-based students was thought worthy of investigation at the same time.

10.3 Summary of Initial Finding

The findings from the initial meeting might be summarised as below:

- Inclusion of the different student support facilities such as Student Services, Careers Service and the Counselling Service in induction to strengthen the student awareness of their existence and the role that they can play;
- Involve the Careers Service to assist students in obtaining relevant work experience;
- Raise the profile of the social side of Halls of Residence with the view to increased student involvement;
- Examine the use of mentors in an effort to increase a student's sense of belonging;
- Look at assessment methods and in particular examinations in an effort to align coursework and examination successes;
- Examine links between different student problems other than those that directly influencing retention;
- Raise the profile of the higher retention rates of home-based students with a possible view to increased efforts to recruit these types of students.

10.4 Follow-up

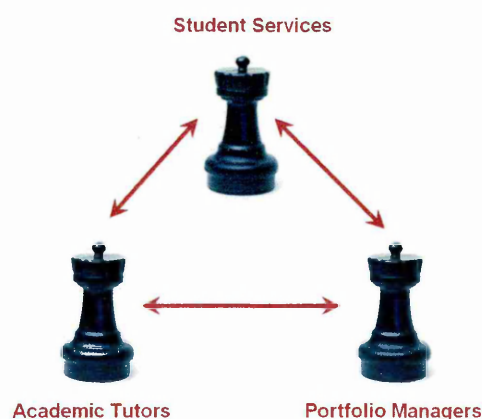
It was decided that the minutes would be circulated to attendees and other members of the Retention Committee by email for comment. Follow-up telephone calls and emails might need to be made to clarify issues. These comments would be organised and a final report compiled by the author. In addition, a member of the Student Services staff should be identified and interviewed as well as other interested parties.

A member of the Student Services staff was identified and informally interviewed along with a couple of other interested members of the faculty. The research findings were presented to them along with the initial recommendations from the Focus Group Meeting. These documents formed a basis for discussion.

Chapter 10
1000 Group Recommendations

It was found that Student Services were very keen to get involved at an early stage, both to increase student awareness of their existence and also as a means of assisting retention rather than as a last resort when a student had got to the stage of desperation. However, if Student Services were to be used much more frequently by students it would necessarily place increased pressure on the staff to respond to student needs. The issue of tutor advice from Student Services was also discussed and felt to be a very profitable way forward.

The Faculty of ACES is divided into a number of portfolios. It has two forms of student support, the Year Tutors and a Team of Portfolio Managers (non-academic staff) who deal with the day to day issues and as a point of initial contact for students. A portfolio is a group of associated courses drawn together for administrative purposes. These are looked upon as twin towers. SHU Student Services is regarded with a high degree of respect but is not used as frequently as it might be. It is proposed that this should become a far more integrated part of the structure. This introduces a third tower. The triangle thus formed would be of greater strength, each tower supporting each other but offering different sorts of expertise. It is felt that much more transparency and cooperation is needed between the different towers to the better good of the students that they are attempting to support. This is shown schematically below:



This structure exists at the moment but is not functioning as efficiently as it might. Each tower should be looked upon as a specific area of expertise in the wholeness of the student experience. Each should exist as part of the whole and not as separate entities with the passing of information and advice as required.

Chapter 10
Focus Group Recommendations

A suggestion was made about the role of work based activities and mentoring. It was felt that it should be much more structured and integrated into the student experience. One way to do that might be by making it "Credit Based". That is, making it worth something towards their degree, perhaps twenty credits, equal to one final year elective module. First and second year students could then be encouraged to pursue some such activities throughout their course and asked to reflect upon them and be assessed in their final year. This could cover things like relevant part-time work, mentoring and voluntary activities such as those set up by the Hallam Volunteers¹. This would help tackle the problem of lack of relevant work experience, self confidence and perhaps reduce the distractions affecting the students. It would increase their commitments but might help to reduce outside commitments that have been seen to have a negative effect upon retention.

The issue of the quality of teaching was raised as a significant problem by some students. Though this was not part of the original study it was thought that it was worthy of further investigation. Since there are two principal aspects to this, those of staff subject knowledge and staff delivery style. Indeed different styles of teaching were thought to benefit different types of environment and types of student. Student perceptions of teaching might be looked upon as modular, in that they are considered in the specific context of the module being taught, but not as a whole across their entire course. It was thought possible that increased awareness of what is being taught in other modules might be beneficial to tutors and might reduce conflicts of teaching method and content which students undoubtedly pick up.

In the correspondence that ensued after the initial meeting the subject of the younger siblings, particular of larger families came up a number of times. Informal discussion had taken place with mothers of primary school children. It was thought to be a well known fact that younger siblings don't do as well as the eldest but whether this was anecdotal or factual was questioned, "within-family sibling comparisons reveal that first born children generally outperform their younger siblings" (Conley et al, n.d.) The idea of whether parents are more relaxed with younger children and hence less 'pushy' might be a cause with the consequence of increased risk of attrition.

¹ Hallam Volunteers is a scheme run from in the university to involve students in voluntary work

10.5 Recommendations

10.5.1 Problem Areas

Before outlining the recommendations the problem areas defined in section 9.5.4 have been restated for ease of reference together with problem number to aid cross-referencing. These problem areas were established as strong antecedents to the considered consequence of "Considered leaving/changing at some stage".

Antecedents

Strongest (from at least two rules)

1. Under 21
2. 2nd Year, 1st Year
3. Non-local
4. From a large family (3 or more children)
5. Younger sibling
6. Halls of residence
7. Have no work experience or only part-time
8. Have no relevant work experience
9. *Have sometimes found the course stressful*
10. *Have had stress problems*
11. *Experience too many distractions that affect ability to work*
12. *Examinations performance has been below expectations*
13. *Bad at managing finances*
14. *Course different from expectation*
15. *Lack self confidence*
16. *Outside commitments high*

Weaker (from only one rule)

17. *Behind with work*
18. *Personal circumstances changed*
19. *Induction didn't help to feel more comfortable*
20. *Criticism disliked*
21. *Taught badly*

Consequence

Considered leaving/changing at some stage

Figure 2

10.5.2 Suggested Recommendations

The recommendations that follow were compiled by the author in consultation with the focus group (Objective 5). These recommendations have been drawn up in an attempt to address the specific identified problems summarised in Figure 2 above and are cross-referenced in the 'Problem Area' column. The recommendations have been specifically designed to attempt to address a number of the problem areas.

As a point of clarification, let us consider the first recommendation below. The Faculty of ACES has planned an Extended Induction Course (EIC), initially for new first year students. The problem areas cross-referenced in the right-hand column below are the ones that the author believes are most likely to be addressed by this extended induction course. The EIC is aimed at first year student, most of which will be under 21 (problems 1 and 2). The content of the

EIC might cover student finance (point 13) and issues that can cause students to get behind with their work (points 16, 17 and 18). It will also help them better understand the course of study that they are enrolled on (point 14). A primary intention of such the EIC will be to improve what is currently on offer (point 19). It is intended that after ratification by the focus group the recommendations will be circulated for wider discussion and possible implementation.

Recommendation	Problem Areas Addressing
1 Examine the proposed Extended Induction Course with the view to the inclusion of Student Services, Careers Service and the Counselling Service to strengthen the student awareness of their existence and the role that they can play	1, 2, 13, 14, 16, 17, 18, 19
2 Involve the Careers Service to assist students in obtaining relevant work experience	7, 8
3 Involve the Counselling Service at an early stage to encourage students to make more use of it in an effort to deal with stress issues	9, 10, 15, 18
4 Strengthen the relationship between Student Services, Portfolio Managers and Academic Tutors to enable a united team of student support	All
5 Raise the profile of the social side of Halls of Residence with the view to increased student involvement. This should be raised at a pan-university level and the Student Union	1, 2, 6
6 Examine the use of mentors, second years for first year students, in an effort to increase a student's sense of belonging. This could be extended in Halls of Residence by the use of 'Senior Students'	1, 2, 15
7 Look at assessment methods and in particular examinations in an effort to align coursework and examination successes	12
8 Examine links between different student problems other than those that directly influence retention, such as socio-economic issues. This might have an effect upon widening participation	All
9 Raise the profile of the higher retention rates of home-based students with a possible view to increased efforts to recruit these type of students	3
10 Examine ways that 'credit based' paid and voluntary work can be integrated into the curricula of courses	7, 8
11 Research student perception of teaching with specific reference to subject knowledge and delivery competence	20, 21
12 Research the issue of younger siblings, particularly from larger families (at least three children) with a view to establishing a firm relationship with great attrition or otherwise	4, 5
13 Conduct telephone interviews with students that have stopped attending in an effort to retain the student and/or discover more about leaving issues	All

Chapter 11: Reflective Summary

11.1 Preamble

In this chapter the author will attempt to evaluate the successes or otherwise of the study, outlining where things might have been done differently with hindsight. It will also define the areas of difficulty and possible compromises that might have been made. It will inevitably discover weaknesses and missed opportunities as well as some blind alleys where time was spent fruitlessly. The evaluation will be tackled by taking each part of the research and examining the issues encountered there.

During the study, great efforts were made to ensure that the direction of the research did not drift, but remained focused on the primary objectives as set out in Section 2.2. The Aim and Objectives have been restated below for easy reference:

Aim

The aim of this study is to explore the issues that affect SHU ITPA computing students and contribute to dropping out, with the view to possible tutor intervention to improve retention rates. It is intended to test the value of using Data Mining techniques in higher education using the vehicle of student retention.

Objectives

1. Identify from literature and secondary qualitative research, the student issues that contribute to dropping out;
2. Investigate quantitatively, by means of a questionnaire, how widespread these issues are;
3. Explore, by means of Data Mining techniques, the interrelationships between the issues;
4. Validate the results of the quantitative research within the ITPA by means of structured interviews;
5. Identify by the use of a ITPA tutor focus group, where tutor intervention can be used to help alleviate these issues;
6. In the light of the findings compile a list of recommendations on how data mining techniques could be further developed.

11.2 Focus of the Research

During the course of the study a number of factors have changed that have a direct impact on the whole area of student retention. The first issue is that of the decline in popularity of computer related courses. The UCAS statistics for Computer Science, as of 15th January 2004, showed a national reduction of

20.1% for undergraduate admission, though the overall number of applications for all subjects was up by 3.8%. However, this trend appears to be slowing down. As of 24th March 2006 the number of applicants for undergraduate courses was down by 9.9%. (UCAS, 2006b) Nevertheless it is putting university computing departments under greater pressure to keep numbers up with a consequential change in the student profile. Student Retention is of vital importance. Financially, it is cheaper to retain a student than to recruit another.

A second change will take effect in the academic year 2006-7. This academic year sees a large increase in student fees to £3,000 in most universities. Though student loans are to be increased in order to meet the fee, it will put new students in much greater debt on leaving university. Indeed it might result in an even greater decline in students wishing to study computer related subjects and the consequential pressure to retain students. All this is going on in a climate of widening and increasing participation.

11.3 Previous Research

There is a wealth of literature concerning student retention in the public domain, however since the study scene is changing, this literature needs to be read and interpreted with caution. Any literature that was written prior to 1998 will not have factored in student debt to any great extent since tuition fees were not introduced until that year though the phasing out of grants and the introduction of student loans had been going on for some time previous to this. The major problem encountered was which literature to include and which to leave out. Reviewing literature from Further Education rather than Higher Education though of relevance, had a number of specific differences that needed to be appreciated. Firstly many of the students were younger and almost exclusively home based. They also covered students from a wider educational base.

The literature review had to be topped up several times and in some cases substantially rewritten during the course of the research. However, some contemporary authors such as Mantz Yorke (1999) have grasped the cost issue, devoting a significant amount of space to it in his book 'Leaving Early'. (Yorke, 1999).

11.4 What is New in this Research

The author, having a background in data analysis and in particular Data Mining was keen to apply his skills in an educational context. Having taught in schools for many years prior to becoming a university lecturer and having recently put his own three children through the university system gave the author a greater incentive to examine the issue of student retention.

There is almost a complete absence of literature concerning Data Mining applied to Higher Education so it was very difficult to put the research into context. However, once it was realised that there were two major dimensions to students, their demographic characteristics (e.g. age, gender and location) and the problems they encounter, the analysis seemed to open up. A university can be likened to a supermarket, the shoppers being the students and the products the problems they encounter (e.g. stress and distractions). This made the form of data collection relatively easy. The analysis of the data would be like the supermarket analysing the buying patterns of the different demographic groups of shoppers.

He believes that the choice of Data Mining as an analysis tool was a sound one that has helped gain a new insight into the domain of student retention. This further demonstrates the value of data mining within higher education.

11.5 Rationale of Methodological Approaches

It is easy to say that universities are not adapting to the changing profile of their students. Many of the institutions have been around for centuries and successfully educated many millions of people in their time. However, society has substantially changed in the last fifty years with more people having the opportunity to attend university than was ever predicted. Now with widening and increasing participation initiatives this is set to increase still more. However, every change presents different challenges and universities need to be ready to adapt their culture to meet these challenges. Many retention initiatives have focused on how the university might assist (induct) students into the general ethos of university life. This is alright as far as it goes, but there are definite signs that this might not be the best course of action. Students from minority groups, lower socio-economic groups, different cultures, students with special needs and the like are beginning to force universities to think the impossible and considering changing to meet these differing student issues.

11.6 Initial Interviews

The initial interviews were very profitable, with the students responding positively. As with any qualitative study it is difficult to know whether you have interviewed sufficient students to get a feel for the issues, or even that you interviewed the right students. However, the issues that were raised largely fitted with the findings from the literature search. A few issues, such as finance and managing finances had become of a higher importance, but generally there were no surprises. On reflection the author believes that if this exercise was to be conducted again he would select more students (say three from each year of study) together with a sample of post-leaving interviews into order to establish the significance of the issues at an earlier stage. However, undertaking the task of post-leaving telephone interviews after the survey had been conducted has given some limited, but positive credence to the findings. (See Appendix K)

11.7 Devising the Questionnaire

The difficulty wasn't what to put into the questionnaire but more what to leave out. The interviews revealed a wealth of issues to investigate but if they had all been put on an online questionnaire it would have been too large to administer. It was felt that to keep the questionnaire to a manageable size would give a greater incentive for students to complete it.

It is believed that the use of a five point scale (1 to 5) for the student issues (see Appendix E, Part 2) was a successful exercise but it did present the author with some significant issues of how to group the responses. For Data Mining purposes the problems needed to be coded as 'Yes' or 'No', furthermore they needed to be coded in such a way that the response of 'Yes' meant that there was a problem. For this reason some of the questions needed to be negated. (See Section 7.4) In addition points 1 and 2 (Strongly Disagree and Tend to Disagree) needed to be grouped together as did 4 and 5 (Tend to Agree and Strongly Agree).

A problem was encountered when deciding where to put a response of 3 (Undecided). If they were placed with the 'Yes' responses this would tend to overestimate the problem. However, if they were placed with the no responses it might result in underestimating the strength of the evidence found. The author finally decided to place them with the negative response (4 and 5) so as to allow the true 'Yes' responses to speak for themselves.

Care was taken to ensure that the responses to the demographic questions (See Appendix E, Part 1) were compatible with other surveys so that cross analyses might be made. For this reason commonly used categories were used for such things as socio-economic group and ethnic group. (See Appendix C) However, one small mistake was made that had a strong impact on the study. Instead of asking the student their actual age, they were asked to choose their age group. This turned out to be too restrictive. The groups used were inappropriate for the type of study being conducted. Virtually all the students (more than 91%) fell into the first two groups, under 21 and 21 to 25, with 56% of the student in the 21 to 25 age group alone. This is an important consideration for future research.

11.8 A Methodology to Use

The author believes that the chosen methodology, CRISP-DM was successfully implemented. It did put constraints on the analysis in that it forced the analyst to go through a rigid set of rules. However, it did yield dividends and a successful outcome was achieved. The use of CRISP-DM rather than its popular counterpart SEMMA was justified in that the vendors of the chosen Data Mining software SPSS Clementine® were partners in the creation of the CRISP-DM methodology. A further justification was that the author was already familiar with it.

11.9 Mining the Data

Much more could have been made of mining the data. The author restricted the analysis to finding relationships that could usefully predict student attrition. However, there is much in the data that is yet to be discovered. This will be discussed further in the next chapter.

There was a great deal of conformity between the different Data Mining techniques used. A number of the findings were of no surprise but some, like younger sibling of large families (at least 3 children) were surprising. Anecdotally it has been known for generations that younger siblings are less likely to do well than older siblings (Conley et al, n.d.), however, putting this in a student retention context needs further investigation. The author was surprised how relatively unimportant financial matters were in the factors that affect student attrition. Managing finances just about scraped into the ratings, but student debt didn't. Again this might be the focus of future research.

11.10 Post-Enquiry Interviews

The author believes that the effort required to contact and conduct these interviews was time well spent, even though the sample was small and not necessarily representative of the students who left. Indeed he believes that it is a process that should be adopted as a normal matter of course when students are believed to have stopped attending university for whatever reason. The interview schedule would need to be revised in the light of experience but it is believed that the concept is sound and revealing.

11.11 Focus-Group Recommendations

Using a pre-constituted forum FARG as a means of selecting a Focus Group might be considered to be rather limiting. However, it did ensure that the members of staff who volunteered were dedicated to the issues of Student Retention. Allowing others to discuss the issues enabled a wider discussion. It is intended that the recommendations from the Focus Group should be put to FARG for ratification at the start of the next academic year.

11.12 Postscript

In general, though there have been some problems and mistakes made, the final outcome has been positive. The author believes that some real progress and insights have been made in the subject area that are beneficial to both the university and the students themselves. The impact that these might have in the longer term and on other institutions is yet to be realised.

Chapter 12: Data Mining Recommendations

12.1 Introduction

This chapter seeks to realise Objective 6, 'In the light of the findings compile a list of recommendations on how the use of data mining techniques could be further developed.'

"Higher education will find larger and wider applications for data mining than its counterpart in the business sector, because higher education institutions carry three (3) duties that are data mining intensive: scientific research that relates to the creation of knowledge, teaching that concerns with the transmission of knowledge, and institutional research that pertains to the use of knowledge for decision making." (Luan, 2002:3)

12.1.1 Online Data Capture

The study used an online questionnaire sited on an external internet web-server to collect the data. The data was stored real-time in a SQL-Server[®] relational database specifically designed for the purpose. This gave the advantages of being able to cover a larger population and not having to input the data manually. However, it did have the disadvantage of the lack of personal contact. The author believes that the advantages of this form of data capture significantly outweigh its disadvantages.

The questionnaire captured two types of data, the student demographics such as age and gender and responses to attitude issues on a five-point attitude scale. The demographic data collection was by the means of a fixed list stored in a database table with a default value shown that reflects the most likely answer to the question. Though this method did force the student to choose from a predetermined list, it did ensure that the collected data always had responses in the permitted range and eliminated the likelihood of missing values. The attitude responses were to pre-worded statements and all used the same scale. This turned out to be very restrictive in that much more care needed to be taken in the wording of the statements to convey the meaning. However, using the same scale did have the effect of reducing the complexity of the data entry form.

On reflection, the author recommends this form of data capture over other means. Although a great deal more effort was needed in the early stages to compile the questionnaire and code the entry form, this was more than compensated for in the ease of speed of data organisation ready for analysis.

12.1.2 Use of Data Mining

The author believes that data mining techniques have proved to be an excellent means of analysing data of this type. His use of three of the data mining techniques, Decision Trees, Rule Induction and Clustering were found to be highly effective and, as stated above, established the primary issues that effect student retention, corresponding well to findings from previous research using standard statistical techniques.

There are a number of other data mining techniques such as Linear Regression and Logistic Regression that could also be used in future research. (Dunham, 2003:80) Neural Networks were found to be useful in this study but yielded a lower level of confidence in the rules it discovered. However, with a much larger data sample the "NNs [neural networks] are more robust than DTs [decision trees] in noisy environments", that is when the data contains such things as missing values or of out of range values. (Dunham, 2003:105)

12.1.3 Required Skills Level

In any research, the skills base of the researchers needs to be assessed and the requisite skills acquired. Though data mining is a complex, extensive field the author believes that the acquisition of the necessary skills is no greater than training the researcher in the standard statistical techniques. Indeed, because the techniques have a number default values associated with them they may be learned and understood more quickly. The author's experience with undergraduate students is that the key skills and concepts can be learned and assimilated in a few weeks.

12.2 Further Work Involving Data Mining

There are a number of areas where additional work might be conducted. Some have already been mooted in the text, particularly in the focus group recommendations. This further work can be defined as falling into four principal areas:

1. Further exploration of the collected data;
2. Extending the retention study;
3. Data mining other available data;
4. Possible new studies involving data mining.

12.3 Further Exploration of the Collected Data

The author has established a number of inter-relationships in the data but restricted the number of factors to three. (See Section 9.4.1) The data could be re-examined to establish a greater dependency between the variables by increasing the number of factors to five or more by the use of MBA.

The original dataset was collected for the purpose of examining interrelationships in the data that may lead to student attrition. This is a narrow scope and has necessarily missed much of the richness of the collected data. There may well be many interesting associations in the data that might indeed be of value. Since data mining can be used in an unsupervised way (see Section 4.2.5 above) then it would be restrictive to put barriers on this exploration. However, some connections, such as socio-economic group, stress and financial matters might be beneficial. Yorke (1999:48) suggests that there is a strong relationship between students who consider themselves to be working class and financial problems. Indeed, with the prospect of significantly more debt after the introduction of increased student fees, it is likely to be this type of student that is worst hit. Their parents are less likely to be able to assist them financially than middle class parents.

In short then, three studies suitable for data mining analysis are proposed:

1. Re-examine the data to look for further inter-relationships between the established retention issues;
2. Examine links between different student problems other than those that directly influence retention, such as socio-economic issues. This might have an effect upon widening participation;
3. Explore the data to find any additional relationships that are worthy of note.

12.4 Extending the Retention Study

12.4.1 Preamble

This study was conceived in order to demonstrate the power of data mining techniques in higher education using the vehicle of student retention issues in the 'New Universities'. The author believes that although the study has been successful in its aims it could benefit from a wider and more far-reaching study involving many more students in more diverse universities. Data mining techniques are scalable to whatever sample size is available. Replicating the study under such conditions would strengthen the evidence gained and enable more transferability to other higher education institutions of different profiles.

12.4.2 Extending the Study within Computing Students

As a starting point the study could be replicated using a much greater sample of students, still being drawn from computing students. This would require the same or a modified questionnaire being administered to a wider and more diverse group of universities. The analysis of this data could take all the universities collectively to gain a general picture and/or be grouped according to the type of university in order to examine any differences between them.

12.4.3 Extending the Study beyond Computing Students

A further extension to the study might widen the remit to students from disciplines other than computing. Again, if the student's subject of study was asked in the questionnaire, the data could be partitioned to allow comparisons to be made. (Dunham, 2003) This would enable wider conclusions to be drawn and enhance the transferability of the survey.

12.5 Data Mining Other Available Data

12.5.1 Students Records

There is a wealth of data that has already been collected within university faculties and beyond. Much of this data has been explored by conventional means, but there may be scope to look at this again by the use of data mining techniques. One such area is historical student records that include degree classifications. The author has already examined this in conjunction with undergraduate students with the view to predicting student degree classifications from student demographic data and certain academic information. There is definite scope here for further investigation.

In addition, results from coursework could be looked at in conjunction with examination results to establish whether there is a real correlation between them. It could also establish whether anecdotal evidence of examination results in modules being perceived to be lower than coursework marks is indeed true. "Correlation can be used to evaluate the strength of a relationship between two variables". (Dunham, 2003:55)

Two recommended studies involving data mining analysis are proposed:

1. Analyse coursework marks and examination marks to see if there is a strong correlation between them. In addition, determine whether the examination results are significantly lower than the coursework marks;
2. Analyse historical student data to build a Data mining model that can be used to predict student degree classifications.

12.5.2 HESA Data

HESA has collected data concerning student registration and admissions over a numbers of years. This data, if mined could reveal interesting demographic inter-relationships and dependencies. Possible areas for investigation might be:

1. Widening participation
2. Minority groups
3. Socio-economic and university type (Traditional, Red-Brick, Plate-Glass, New)
4. The growth or otherwise of 'home-based' university education

12.6 Possible New Studies Involving Data Mining

There are many additional studies that could be performed as a direct consequence of this study which would require new data to be collected. This might involve looking at the issue of Widening Participation to see if we are indeed reaching out to the under-represented groups in our society and increasing the representation of these students on our courses. This would involve an ongoing exercise over the next three to four years to ensure that demographic data is captured that can be compared on a year by year basis. Presented below are three recommended studies that would lend themselves to analysis by data mining techniques:

1. Research into student perception of teaching with specific reference to subject knowledge and delivery competence with respect to student progression;
2. Research the issue of younger siblings, particularly from larger families (at least three children) with a view to establishing a firm relationship with greater attrition or otherwise;
3. Investigate on a year by year basis the proportion of students from each socio-economic group to establish whether we are widening participation.

Chapter 13: Conclusions

13.1 A Note of Guidance

Student Retention and Data Mining Techniques

The author wishes to present this study as an evaluation of Data Mining Techniques using Student Retention Issues as a vehicle.

The approach of this thesis - demonstrating the potential of data mining as a research technique in Higher Education by applying it to the specific issue of student retention - necessarily leads to a difficulty of presentation. The substantive research area of student retention has had to be considered in appropriate detail, which can at times mean that its function as a vehicle to demonstrate data mining disappears from view. It is therefore important here to assert the prime argument.

13.2 Summary

"Students tended to give multiple reasons for withdrawal, the primary reason often being difficult to determine." (Yorke, 1999:25)

This seems a suitable remark to start the summary. In short, students are generally not fully aware of their reasons for leaving. Indeed it is likely that it is a combination of reasons. The way that these factors interact will determine whether the student will make that decision to leave, or not. This study has attempted to unpack these multiple forces and examine their interaction. It has tried to identify the primary issues, both demographic and problem based to arrive at the major at-risk factors. These were summarised at the end of Chapter 9 (Figure 2); a précis is repeated below.

Antecedents

Strongest (from at least two rules)

- 1 Under 21
- 2 2nd Year, 1st Year
- 3 Non-local
- 4 From a large family (3 or more children)
- 5 Younger sibling
- 6 Halls of residence
- 7 Have no work experience or only part-time
- 8 Have no relevant work experience
- 9 Have sometimes found the course stressful
- 10 Have had stress problems
- 11 Experience too many distractions that affect ability to work
- 12 Examinations performance has been below expectations
- 13 Bad at managing finances
- 14 Course different from expectation
- 15 Lack self confidence
- 16 Outside commitments high

Weaker (from only one rule)

- 17 Behind with work
- 18 Personal circumstances changed
- 19 Induction didn't help to feel more comfortable
- 20 Criticism disliked
- 21 Taught badly

Consequence

Considered leaving/changing at some stage

Figure 2 (Reproduced from Section 10.5)

Chapter 13

A number of the above factors have been raised in previous literature. For instance, all those in italics above were identified, at least in part by Yorke (1999:38). This is encouraging as it shows that the survey mechanism and data mining analysis does conform to previous research in a positive way. That is, eleven out of the twenty-one factors had previously been identified as major at-risk factors.

However, there are a number of strong connections that have not appeared in previous literature, for instance the younger siblings of large (greater than 3 children) families. This seems significant and worthy of further research. As stated before, anecdotally this is of no great surprise.

In a subject like computing, which is thought to be vocational, the need for relevant work experience might be crucial. To study the subject without any opportunity to see or put it into operation might be looked upon as a demotivating factor. However, having little or no work experience, regardless of it being relevant seems worthy of mention. This might suggest that students should always take time out to work, prior to going to university as a matter of course!

It is not surprising that students in the first year are more at risk, since they are generally younger and less mature (Under 21). However, a greater factor will be the number of these that have already dropped out before their final year hence making the final year students appear to be less vulnerable.

If these factors are genuinely correct then the earlier they can be identified the better chance we might have of avoiding students leaving.

13.3 Use of Data Mining

The author believes that he has demonstrated that data mining techniques can be of great value in the understanding of student data through the vehicle of student retention issues. The techniques can be used descriptively to produce graphs and charts to get an initial picture of the data, as well as analytically. They have a great strength over traditional statistical techniques in that they can be used inductively without any preconception of what is likely to be found. They can also be used effectively by non-statisticians. This study has demonstrated that data mining techniques have found most if not all of the issues established in previous research effectively and attempted to find the strongest relationships between them.

13.4 Summary of Data Mining Recommendations

13.4.1 Introduction

In chapter 12 the author has made a number of recommendations for the possible use of data mining techniques within higher education. What follows is a summary of the recommendations made in Section 12.1 to 12.5 above.

13.4.2 Data Collection Methods

Much more use should be made of existing data that have been accumulating in all universities for a number of years. Unfortunately much of this data is stored in a series of heterogeneous data sources. Once the data has been gathered together this will represent a much greater resource.

Where new data needs to be collected from questionnaires then the author recommends the use of online questions made available over the World Wide Web. The collected data should then be stored in a resilient relational database such as Microsoft SQL- Server[®].

13.4.3 Data Analysis Methods

As stated above the author believes the use of data mining techniques in this study have proved to be a highly successful means of data analysis. Below are outlined where he believes that these can be used to great effect.

13.4.4 Further Exploration of the Collected Data

1. Examine links between different student problems other than those that directly influence retention, such as socio-economic issues. This might have an effect upon widening participation;
2. Explore the data to find any additional relationships that are worthy of note.

13.4.5 Extending the Study

Repeat the study using the same or a modified questionnaire to a much greater sample of students situated in a wider group of universities and other subject areas. This might involve looking at minority issues such as females in computing and the high incidence of students of Asian descent.

13.4.6 Data Mining Other Available Data

1. Analyse coursework marks and examination marks to see if there is a strong correlation between them. In addition, determine whether the examination results are significantly lower than the coursework marks;
2. Analyse historical student data to build a Data mining model that can be used to predict student degree classifications.

13.4.7 Possible New Studies Involving Data Mining

1. Research student perception of teaching with specific reference to subject knowledge and delivery competence with respect to student progression;
2. Research the issue of younger siblings, particularly from larger families (at least three children) with a view to establishing a firm relationship with greater attrition or otherwise;

Investigate on a year by year basis the proportion of students from each socio-economic group to establish whether we are widening participation.

13.5 *Postscript*

Looking back at the recommendations of this study, there is much that is still to be done. However, after conducting this study the author feels enabled and empowered to continue researching the use of data mining techniques in the context of higher education.

This student retention study draws upon previous research and adds to it the new dimension of Data mining. The author's background in this field has enabled this developing discipline to be used in a new and exciting way. Undoubtedly there are many other areas of higher education that could benefit from the use of Data mining techniques, some of which have been mentioned in the recommendations above. In the course of further research it is expected that the author will draw upon this experience and take it forward, possibly working with like minded colleagues and postgraduate students either at Masters level or Doctorate level.

This study has not been exhaustive, but has been thorough in the areas that have been explored. A number of new insights have been discovered that have taken research forward in a new and exciting direction.

In conclusion it must be pointed out that this study was conducted entirely with students from Computer related courses. It would be unwise to assume that the retention findings are fully transferable to other disciplines.

"The patterns of influence on withdrawal differed from one Academic Subject Category to another. It seems probable that some of the differences relate to cultures of the respective disciplines". (Yorke, 1999:55)

References

- Aim Higher (2006) *Thinking about your future? Aimhigher! 'Part-time Work'*, HEFCE [online] http://www.aimhigher.ac.uk/student_finance/part_time_work.cfm [1st August 2006]
- Airms (2004) *Guide of Standards for Marketing and Social Research* [online] http://www.airms.org/pages/an/focus_groups.htm [1st August 2006]
- Archer L (2002) *A Question of Motives*, Times Higher Education Supplement, 31st May 2002
- Ashby A (2004) *Monitoring student retention in the Open University: definition, measurement, interpretation and action*, Open Learning 19[1]:65-77, February 2004, Milton Keynes: The Open University
- Astin A (1993) *What Matters in College? Four Critical Years revisited*, San Francisco: Jossey-Bass
- Bean J & Metzner B (1985) *A conceptual model of nontraditional undergraduate student attrition*, Review of Educational Review, 55:485-540
- Beatty-Guenter P (1994), *Sorting, Supporting, Connecting and Transforming: Retention strategies at Community Colleges*, Community College Journal of Research 18[2]:113-129
- Berger J (2001-2) Understanding the Organizational nature of Student persistence: Empirically-based Recommendations for Practice, Journal of College Student Retention: Research, Theory and Practice 3[1]: 3-21, Amityville: Baywood Publishing Company, Inc
- Berry J & Linoff G (2000) *Mastering Data Mining: The Art and Science of Customer Relationship Management*, New York: John Wiley & Sons inc
- Berry M, Linoff G (2004), *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*, New York: John Wiley & Sons inc
- Bertino E, Catania B, Kotsifakos E, Maddalena A, Ntoutsis I and Theodoridis Y (2004), *PBMS Querying and Storage Issues*, Panda Technical Report Series, 20th February 2004, PANDA Consortium, [online] <http://dke.cti.gr/panda/publications/TR-2004-01.pdf> - [1st August 2006]
- Braxton J & Lien L (2000) "The Viability of Academic Integration as a Central Construct in Tinto's Interactionist Theory of College Student Departure", Nashville: Vanderbilt University Press
- Burch A (2003) *Missed Opportunities? Combining Quantitative and Qualitative Methodologies, Mixed Methods Seminar II*, 2nd June 2003, London: Institute of Education [online] <http://www.mlwin.com/team/jonmms2.html> [1st August 2006]
- Cabrera A, Castaneda M, Nora A & Hengstler D (1992) "The Convergence between Two Theories of College Persistence", Journal of Higher Education 70[2]:134-60
- Callender C (2001) Changing student finances in higher education: Policy contradictions under New Labour, Journal of Widening Participation & Lifelong Learning 3[2]
- CEM (n.d.) Qualitative Research, The College of Emergency Medicine (CEM) [online] http://www.emergencymed.org.uk/CEM/Research/technical_guide/qual.htm [1st August 2006]
- Chickering A (1969), *Education and Identity*, San Francisco: Jossey-Bass
- Cloonan M (2004) Notions of Flexibility in UK Higher Education: Core and Periphery Re-Visited? Higher Education Quarterly 58[2,3]:176 April 2004
- Cohen L, Manion L & Morrison K (2000) *Research Methods in Education*, 5th Edn, London: RoutledgeFalmer
- Coldeway D (1986) Learner Characteristics and Success, in: I Mugridge & D Kaufman (Eds) *Distance Education in Canada* (London, Croom Helm), 81-93
- Conley D, Pfeiffer K & Velaz M (n.d.) 'Explaining Sibling Differences in Achievement and Behavioral Outcomes: The Importance of Within- and Between-Family Factors', New York University [online] http://homepages.nyu.edu/~dc66/pdf/sibs_development.pdf [1st August 2006]
- Connor H, Dewson S with Tyers C, Eccles J, Regan J & Aston J (2001) Social Class Participation, DFEE Research Report RR267
- CRISP-DM (2000), *Step-by-step Data Mining Guide*, NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringen en Bank Groep B.V (The Netherlands). Copyright® 1999, 2000
- Dearing Report (1997) *Higher Education in the Learning Society - The Report of the National Committee of Inquiry into Higher Education*, chaired by Sir Ron (now Lord) Dearing, London: HMSO

- DES (1992) *Statistical Bulletin: Leaving Rates Among First Year Degree Students in English Polytechnics and Colleges*, London: HMSO
- Dunham M (2003) *Data Mining: Introductory and Advanced Topics*, New Jersey: Pearson Education inc
- Etimage (1997) *Neural Networks with Java: Neural Net Overview*, Jochen Fröhlich [online] <http://www.etimage.com/java/appletNN/NNtyper/e-11-text.html> [1st August 2006]
- Exchange (2002) Exchange 1:13 2002, National Coordination Team (NCT), Open University
- Fisk D (2006) Beer and Nappies -- A Data Mining Urban Legend [online] <http://web.onetel.net.uk/~hibou/Beer%20and%20Nappies.html> [1st August 2006]
- Forsyth A & Furlong A (2000) *Socio-economic disadvantage & access to higher education*, Bristol: Policy Press & Joseph Rowntree Foundation
- Hartigan J & Wong, M (1979) *Algorithm AS136, a K-means clustering algorithm*, *Applied Statistics*, 28:100-108
- HEFCE (2002) *Information on Quality and Standards in Higher Education*, Report 02/15, Final report of the Task Group (HEFCE 2002/52)
- House of Commons (2001) *Improving Student Achievement in English Higher Education*, London: The Stationary Office
- HCSCEE (2001) House of Commons Select Committee on Education and Employment, *Higher Education: Student Retention Sixth Report H.C. (2000-01)*, 14th March 2001, London: The Stationary Office
- Hyperdictionary (2004) [online] <http://www.hyperdictionary.com/computing/data+mining> [1st August 2006]
- Jefferson (2002) *Ideas to Encourage Student Retention*, Kentucky: Jefferson Community College
- John G & Langley P (1996) *Static Versus Dynamic Sampling for Data Mining*, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*: 367-370, Menlo Park, California: AAAI Press
- Johnes J (1990) *Determinants of Student Wastage in Higher Education*, *Studies in Higher Education* 15[1]:87-99
- Johnson G & Buck G (1995) *Students' Personal and Academic Attributions of University Withdrawal*, *Canadian Journal of Higher Education*, 25[2]:53-77
- Johnston V (2001) *Developing Strategies to Improve Student Retention: Reflections from the Work of Napier University's Student Retention Project*, SRHE Conference paper [online] http://www.napier.ac.uk/qes/studentretentionproject/Documents/SRHE_2001_Cambridge.doc [1st August 2006]
- KDD (1995) *The 1995 Conference in Knowledge Discovery in Databases (KDD)*
- Kennedy H (1997) *Learning Works: Widening Participation in Further Education*, Coventry: FEFC
- Kuonen D (2005) *Is Data Mining for Gold "Statistical déjà vu"?* Statoo Consulting: Lausanne, Switzerland [online] <http://www.crm2day.com/library/EEpZEIkAkyJMPQLSal.php> [1st August 2006]
- Kul G & Love P (2000) 'A Cultural Perspective on Student Departure', in J Braxton (ed) *Reworking the Student Departure Puzzle*, Pp 196-212, Nashville: Vanderbilt University Press
- Lall M, Morley L & Gillborn D (2003) *Widening participation in Higher Education Research Report 11*, London: University of London Institute of Education
- Long M, Carpenter P & Hayden M (1995) *Graduating from Higher Education*, Canberra: Australian Government Publishing Service
- Luan J (2004) *Data Mining Applications in Higher Education*, SPSS Incorporated Executive Report, SPSS: Chicago
- Luan J (2002) *Data Mining and Knowledge Management in Higher Education - Potential Applications* Presentation at AIR Forum, Toronto, Canada [online] http://www.cabrillo.edu/services/pro/oir_reports/DM_KM2002AIR.pdf [2nd December 2006]
- MacQueen, J (1967) Some methods for classification and analysis of multivariate observations, in Le Cam L & Neyman J eds, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281-297, Berkeley, California: University of California Press
- Marcinowicz L, Borzuchowska A, Grebowski R (2002) *Methodologic difficulties in measuring patient satisfaction* 55 Suppl 1:335-340 - [online] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=15002265&dopt=Abstract [1st August 2006]

- Martinez P (1995) *Student Retention in further and adult education*, Bristol: FEDA
- Martinez P (1996) *Student Retention: case studies of strategies that work 1[6]* Bristol: FEDA
- McEwen M, Roper L, Bryant D & Langa M (1996) *Incorporating the development of African-American students into psychosocial theories of student development* in F Stage, A Stage, D Hossler & G Anaya (Eds) *College students: The evolving nature of research*:217-226
Needham Heights, Simon & Schuster
- McGivney V (1996) *Staying or Leaving the Course*, Leicester: NIACE
- McInnis C, James R & McNaught C (1995) *First Year on Campus: Diversity in the Initial Experiences of Australian Undergraduates*, Melbourne: Centre for the Study of Higher Education, Melbourne: University of Melbourne
- McInnis C, James R, Polesel J & Teese R (2000a) *Non-Completion in Vocational Education and Training and Higher Education*, Canberra: DEETYA
- McInnis C, James R & Hartley R (2000b) *Trends in First Year Experience in Australian Universities*, Melbourne: Department of employment, Education, Training and Youth Affairs
- McPherson AF and Paterson LJ (1990) *Undergraduate non-completion rates: A comment*, Higher Education, 19:377-383
- Miller, N (2002) 'Designing Questionnaires' in The Handbook for Economics Lecturers, The Higher Education Academy [online] <http://www.economicsnetwork.ac.uk/handbook/> [1st August 2006]
- Morgan D (1997) *Focus Groups as Qualitative Research*, 2nd Edn, Thousand Oaks (USA): Sage
- Moore R (1995) *Retention Rates: Research Project*, Sheffield, Sheffield Hallam University
- Moxley D, Najor-Durack A & Dumbridge C (2001) *Keeping Students in Higher Education*, London: Kogan
- Naisbitt J (1982) *Ten New Directions Transforming our Lives*, New York: Warner
- Napoli A & Wortman P (1997) *Psychosocial factors related to retention and early departure of two-year community college students*, Retrieved 9th August 2001, from Suffolk County Community College, Office of Institutional Research and Assessment [online] <http://sccaix1.sunysuffolk.edu/Web/Central/IT/InstResearch/rhe97.htm> [1st August 2006]
- Oates, T & Jensen D (1997) *The Effects of Training Set Size on Decision Tree Complexity*, Machine Learning: Proceedings of the Fourteenth International Conference: 254-262, San Francisco: Morgan Kaufmann
- Oppenheim A (1992) *Questionnaire Design, Interviewing and Attitude Measurement*, London: Pinter
- Ozga J & Sukhnandan L (1997) *Undergraduate non-completion in higher education in England*. Report 2, Bristol: HEFC
- Padilla R, Trevino J, Conzalez K & Trevino J (1997) *Developing Local Models of Minority Student Success in College*, Journal of College Student development 38[2]: 125-35
- Peelo M & Wareham T eds (2002) *Failing Students in Higher Education*, The Society for Research into Higher Education, Buckingham: Open University Press
- Pendse N (2005) What do all the TLAs and jargon really mean?, The OLAP Report - Last updated on April 1st, 2005 [online] <http://www.olapreport.com/glossary.htm> [1st August 2006]
- Pugsley L (1998) *Throwing Your Brains at it: higher education, markets and choice* in International Studies in Sociology of Education 8[1]:71-92
- Reay D (1998) Always knowing and never being sure: familiar and institutional habituses and higher education choice in Journal of Education Policy 13[4]: 519-529
- Rendon L, Jalomo R & Nora A (2000) 'Theoretical Considerations in the Study of Minority Student Retention in Higher Education', in J Braxton (ed) *Reworking the Student Departure Puzzle*, pp 127-56, Nashville: Vanderbilt University Press
- Rogerson S (1997) Women in IT, ETHcol in the IMIS Journal 7[6] (December 1997)
- Roiger R & Geatz M (2003) *Data Mining: A Tutorial-based Primer*, New Jersey: Pearson Education inc
- SAS (2006a) *Data and Text Mining*, SAS Institute Incorporated [online] <http://www.sas.com/technologies/analytics/datamining/> [1st August 2006]
- SAS (2006b) *Enterprise Miner* SAS Institute inc website, [online] <http://www.sas.com> [1st August 2006]
- SAS (2006c) *University of Alabama identifies, mentors at-risk students*, SAS Advanced Student Retention [online] <http://www.sas.com/success/ua.html> [1st August 2006]

- Sheatsley P (1983) "Questionnaire Construction & Item Writing" In Rossi P, Wright J & Anderson A (Eds.), *Handbook of Survey Research: Quantitative Studies in Social Relations* (pp 195-230), New York: Academic Press
- Slack K & Casey L (2002) *The best years of your life? Contrasting the local & the non-local student experience of higher education*, Stoke on Trent: BERA Staffordshire University Press
- SPSS (2003) *Crime Analyst Says Data Mining Puts More Science and Less Fiction into Law Enforcement Decision Making*, SPSS Press Release, Chicago [online] http://www.spss.com/press/template_view.cfm?PR_ID=579 [1st August 2006]
- Stacey M (1969) *Methods in Social Research*, Oxford: Pergamon Press
- Statistica (2006) *Data Mining Techniques*, StatSoft, Inc [online] <http://www.statsoft.com/textbook/stdatmin.html> [1st August 2006]
- Tait J (2004), *The tutor/facilitator role in student retention*, Open Learning 19[1] February 2004, The Open University, UK
- THES (1995a) Times Higher Education Supplement, 22nd December 1995, OU Counsellor
- Thomas E (2002) *Student Retention in Higher Education: The Role of Institutional Habitus*, Journal of Educational Policy 17[4]:423-442
- Tierney W (2000) 'Power, Identity and the Dilemma of College Student departure', in J Braxton (ed) *Reworking the Student Departure Puzzle*, pp 213-34, Nashville: Vanderbilt University Press
- Tight M (1998) *Education, Education, Education! The vision of lifelong learning in the Kennedy, Dearing and Fryer reports*, Oxford Review of Education, 24[4]:473-485
- Tinto V (1993) *Leaving college: rethinking the causes and cures of student attrition*. 2nd ed, Chicago: University of Chicago Press
- UCAS (2002) *Paving the Way Project Report*
- UCAS (2006a), 'Final figures for 2005 entry up by 7.4%', News Release 19th January 2006 [online] <http://www.ucas.com/new/press/news190106.html> [1st August 2006]
- UCAS (2006b), 'Latest UCAS statistics show number of applicants holding up', News Release 27th April 2006 – [online] on <http://www.ucas.com/new/press/news270406/index.html> [1st August 2006]
- University of Hertfordshire (2000) *Structured interviews/semi structured interviews*, Higher Education Corporation, First Published in April 1998 [online] <http://perseus.herts.ac.uk/> [1st August 2006]
- University of Southampton (2006) *Part-time Work*: University of Southampton Website <http://www.soton.ac.uk/study/careerprospects/parttimework.html> [1st August 2006]
- Watterson K (1995) *Intelligent agents, multidimensional analysis tools, and good old database queries all belong in the well-equipped data miner's toolbox*, New York: CMP Media LLC [online] <http://www.byte.com/art/9510/sec8/art8.htm> [1st August 2006]
- Westphal C and Blaxton T (1998) *Data Mining Solutions: Methods and Tools for Solving real-World Problems*, John Wiley & Sons inc: New York
- Wikipedia (2006) The Free Encyclopaedia, on <http://en.wikipedia.org> [1st August 2006]
- Yorke M (1999) *Leaving Early*, London: Falmer
- Yorke M & Longden B (2004) *Retention and Student Success in Higher Education*, The Society for Research into Higher Education, Maidenhead: Open University Press
- Zepke N & Leach L (2005) *Integration and adaptation: Approaches to the student retention and achievement puzzle*, *Active Learning in Higher Education*, The Higher Education Academy & SAGE 6[1]:46-59: New Delhi

Bibliography

- Bean J (1983) *The application of a model of turnover in work organizations to the student attrition process*, Review of Higher Education 60:155-182
- Cabrera A, Castaneda M, Nora A & Hengstler D (1992) *The Convergence between Two Theories of College Persistence*, Journal of Higher Education 70[2]:134-60
- Clementine (), *Clementine User Guide*, SPSS Software
- Davies P (1999) *Student retention in further education: a problem of quality or of student finance?* Further Education Development Agency, Brighton: University of Sussex
- Financial Times - 8th December 1993
- Gilbert S & Auger N (1988) *Student Finances and University Attrition*, Ottawa: Department of Secretary of State of Canada
- Johnes J & Taylor J (1990) *Undergraduate non-completion rates: a reply*, Higher Education 19 385-390
- Johnes J & Taylor J (1991) *Non-completion of a degree course and its effect on the subsequent experience in the labour market*, Studies in Higher Education 16[1]:73-81
- Jones R & Thomas L (2001) *Not 'Just Passing Through': Making Retention Work for Present and Future Learners*, Widening Participation & Lifelong Learning 3[2]: Staffordshire University
- Kinnock M & Ricks M (1993) *Student Retention: Moving from Numbers to Action*, Research in Higher Education 34[1]
- Mallette B & Cabrera A (1991) *Determinants of withdrawal behavior: An exploratory study*, Research in Higher Education, 32[2]: 179-94
- McKeown B, MacDonell A and Bowman C (1993) *The Point of View of the Student in Attrition Research*, Canadian Journal of Higher Education, 23(2), 65-85
- Pascarella E & Terenzini P (1991) *How College Affects Students: Findings and Insights from Twenty Years of Research*, San Francisco: Jossey-Bass
- Project Gold (2000) *Research Methods Glossary* [online] <http://www.bath.ac.uk/lifelong-learning/> [1st August 2006]
- Reay D (2001) *Finding or losing yourself?: working class relationships to education* in Journal of Education Policy 16[4]:333-34
- Spedding T & Gregson M (2000) *Widening participation: A missing curriculum? Some observations on the policy and practice of widening participation in further education in the UK*, Paper presented at British Educational Research Association Annual Conference, 7-9 September 2000, Cardiff: Cardiff University
- Taylor J & Johnes J (1989) *An Evaluation of Performance Indicators Based Upon the First Destination of University Graduates*, Studies in Higher Education, 14[2], 201-217
- Taylor P (2000) *The engagement of minority ethnic groups in higher education: experiences from the UK*, in Thomas E & Cooper M (eds) *Changing the Culture of the Campus: Towards an inclusive higher education*, Stoke on Trent: Staffordshire University Press
- Teese R (2000) *Non-Completion in Vocational Education & Training & Higher Education*, Canberra: DEETYA
- THES (1995b) Times Higher Education Supplement, 27th January 1995
- THES (2002) Times Higher Education Supplement, 13th January 2002
- Tinto V (1975) *Dropout from higher education: A theoretical synthesis of recent research*, Review of Educational Research, 45:89-127
- Walker R (2000) *Indigenous Performance in Western Australia Universities: Reframing Retention and Success*, Canberra: DEETYA
- Woodrow M (2001) *Are access policies and funding arrangements compatible? Access and Retention in Higher Education*, Conference organised by the Centre for Higher Education Research and Information, 18th May 2001, London

Appendix A: Student Interviews

Six students were interviewed and named Student A-F. Each interview was tape recorded and transcribed verbatim before the analysis began. What follows is an extract taken from the interview of Student A.

Interview of Student A

SPEAKER	SPOKEN	NOTES
Interviewer	Thanks very much for agreeing to take part in this interview. I want to assure you that whatever you say will be in strict confidence. You will not be referred to by name only as student A.	Introduction
Student A	Yes, okay, fine.	
Interviewer	The research is looking at why in Britain we have such a high drop-out rate from university, about 20% from undergraduate courses, looking at the factors that cause this non-completion what we can do to try to improve the situation. I have divided it up into six different areas. I'll briefly go through each one of them and it will be for you to interpret, If you wish, what you believe belongs to each one of these.	Reasons for research
Student A	Okay.	
Interviewer	The first one is associated with the course itself, the second is to do with the institution (the university in which you study), the third is to do with the academic support you received, the fourth one is personal things to do with yourself. The fifth is partially to do with yourself in that it is your personal motivation or lack of motivation. The final one is the big one, financial constraints.	Explanation of procedure
Student A	Okay	
Interviewer	What I am going to do, and it's quite open ended is this, I would like you to come up with some ideas as to what problems you might have encountered concerned with the course, the type of course and your expectations of that course from the point at which you first started.	*** Course ***
Student A	Right, to do with courses, I think the entry levels, when you have to get so many points to get in onto that particular course. That was a problem when I first came onto the course. The course itself would have been, is it too difficult? Is it above me kind of? Can I do this? Erm, is the material interesting? The teaching staff on the course as well. I can see that that could come under a different subject area. Will the course lead me into a job that I want? Is the qualification still valid now?	
Interviewer	Okay, anything else you can think of?	Prompt
Student A	Not really, no.	
Interviewer	Okay, thank you very much. You can always look back at a later stage if you think of anything else.	Thanks
Student A	Okay	
Interviewer	The second one is the institution, that about this university, its facilities and, not the academic side of it, and its location within Sheffield etc. So if you would like to sort of go through that ...	*** Institution ***

Student A	I think that the location of the university is important, is it located in too lively an area? Are there too many distractions. Is it situated in too quiet an area so that you can be bored and can't mix with people. Are there facilities of quality? Can I get to a computer if I need to use one? Books as well, I suppose the type of university, is it a 'red-brick' one? I suppose that will have some effect on whether I complete the course or not.	
Interviewer	Do you think 'red-brick' universities perhaps don't give quite as much student support?	Supplementary question
Student A	Yes, yes, yes. I suppose it comes down to confidence as well. I think that in a 'red-brick' university I would be less likely to approach the lecturing staff as much as I do in this university.	
Interviewer	What about the study and support facilities then?	Supplementary question
Student A	The study facilities, the library and the computers; you've got to have PCs to be able to have access to books and you have got to have access to on-line journals as well.	
Interviewer	So looking back a bit now, which one of those would you say were a demotivating influence to you? I believe you are a day student aren't you?	Looking back Supplementary question
Student A	Yes. I suppose the demotivating influence is if the course is too hard and you can't get help from the staff as well, they are two big demotivating influences.	
Interviewer	So you've experienced both problems have you?	Supplementary question
Student A	Yes, I have.	
Interviewer	Anything else of the first two categories?	Looking back
Student A	I think that the resources as well. If you can't get hold of the resources, books etc.	
Interviewer	Thanks. Let's get onto the third category academic then. Now of course there is going to be a bit of an overlap with the institutional but we are particularly looking at now study skills and tutor support, those sort of things.	Thanks *** Academic ***
Student A	My study skills?	Question
Interviewer	Yes, yes, yeh. If you would like to think about that.	Answer
Student A	I think that lack of English ability, not being able to structure essays.	
Interviewer	Was that a particular problem for you?	Supplementary question
Student A	Oh, yes, yes, it was a problem for me.	
Interviewer	One that you have largely overcome?	Supplementary question
Student A	Yes, one that I have largely overcome.	
Interviewer	Could it have caused you to drop-out?	Supplementary question
Student A	Possibly, yes. Learning to structure essay properly, spelling and grammar.	
Interviewer	What about assessment?	Supplementary question
Student A	Assessment, exams are always a problem.	
Interviewer	Was that a problem all through school then?	Supplementary question
Student A	Yes, I would say yes. I have hopefully cracked it on the head by now. Yes, exams have been a problems, coursework's not so bad.	
Interviewer	Did you always attend lectures and tutorials?	Supplementary question
Student A	Yes	
Interviewer	Were there many people who didn't?	Supplementary question

Student A	Erm, there were always the odd few.	
Interviewer	Do you think that that can be a demotivating influence?	Supplementary question
Student A	Yes, I suppose so, you are not getting the familiarity with the lecturer and the material. I think the other thing with lectures and tutorials is that there is a fine balance if you've got work to do and if you need to go to the tutorials. If you got a project to write and it's due in a couple of days then I might have to miss some lectures and tutorials because of the stress and the strains.	
Interviewer	What about getting behind with your work?	Supplementary question
Student A	I have never actually got behind with my work.	
Interviewer	That's not been a problem then?	Supplementary question
Student A	No.	
Interviewer	You think that if you had got behind, it would have caused you a problem?	
Student A	Yes.	
Interviewer	Would you like to pin-point anything in particular that you feel was a problem to you? Any one particular ...	Supplementary question
Student A	Yes, job commitment. I was working part-time as well as doing my degree.	Supplementary question
Interviewer	Does that perhaps come more under personal problems?	Postponement? *** Personal ***
Student A	Yes, erm ... I suppose criticism of lecturers and tutors is I suppose a problem of mine as well.	
Interviewer	Okay, let's have a look at personal things now. These difficulties might range from having difficulty settling in when you first came here to any problems from home, illness, lack of friends, lack of self confidence ..., alcohol problems ...	Personal
Student A	Yes, working part-time, I had a girl friend and it was going up and down a little and I think it contributed to me becoming a generic student. (Failing a module). The pressures that I had were from the work.	
Interviewer	You failed a particular module then?	Supplementary question
Student A	Yes, okay.	
Interviewer	Did you take a course to replace that one then?	Supplementary question
Student A	Yes, I did.	
Interviewer	Illness or accidents, anything like that?	Supplementary question
Student A	Illness, could be, but when ever I have been ill I have always forced myself to come in. The problem with the job commitment and the girl friends was, it didn't affect lectures or tutorials it meant that my coursework was rushed and my exam revision was more rushed.	
Interviewer	Would you like to think of one in particular you feel was perhaps the strongest distraction?	Supplementary question
Student A	I'd go with relationships.	
Interviewer	Did it make you feel low?	Supplementary question
Student A	Yes, it made me feel low, distracted me from my work.	
Interviewer	Right, let's look into the next section which I've broadly called motivational. These are things that drive you forward, or not.	*** Motivational ***
Student A	Okay, if you want to know what drives me it's the fear of failing. I didn't get good A' levels	

	and I'm very much frightened of failing. I feel that I have got something to prove to someone all the time.	
Interviewer	Is that always going to be the case?	Supplementary question
Student A	Yes. I think that I go through my life very highly tensed and stressed. I think that demotivational factors are at home. My mother's very much wanting to talk to me all the time when I'm trying to work. That's quite a problem.	
Interviewer	You've always lived at home?	Supplementary question
Student A	Yes, I have yes.	
Interviewer	Was that a choice or did you have pressure to stay at home?	Supplementary question
Student A	It was a personal choice, because I got bad A' levels I did a foundation year in Computing and Management Sciences at Tapton, Chesterfield College, which got me into university to do a degree, so I really had no choice. I suppose there were financial reasons as well.	
Interviewer	Could you please perhaps consider any one particular thing again that you feel may have been your principal component in lack of motivation or strength?	Looking back Supplementary question
Student A	My strength of motivation is fear of failing. Lack of motivation would possibly be when it gets too hard I tend to switch off.	
Interviewer	Right, okay. Does that happen in lectures?	Supplementary question
Student A	Erm, yes if it's too hard or I already know it I sometimes tend to switch off.	
Interviewer	Finally, and this one that grown a lot over the last few years is finance. There may be some differences since you were a day student. Did you get a smaller loan?	*** Financial ***
Student A	Oh no, I got the same loan.	
Interviewer	Would you like to describe this? I know, because of it's a personal issues you might not want to fully talk about it, but considering the confidentiality of this I would just like to unpick any financial problems that you might have had that could have caused you to drop out, have brought you to the brink at least.	Supplementary question
Student A	I think that this is entwined with the job commitment is financial. I didn't get a student loan in my first year so my hours were up so that I could support myself.	
Interviewer	Did you deliberately not go for a loan?	Supplementary question
Student A	Yes, but when I found out my marks and because generic because I failed a unit I cut my hours, which meant that I had less money coming in but I was still going out on a Friday night with my friends so I had to get a student loan.	
Interviewer	So you'll have a very large debt then?	Supplementary question
Student A	Yes I will. Yes I'm afraid of failing but also because of failing for financial reasons.	
Interviewer	Have you had the support of your parents?	Supplementary question
Student A	My parents have paid my tuition fees only. I have to pay board at home, but my parents believe that it is my debt and it's what I want to do, so it is all me.	
Interviewer	Would you like to perhaps come up with the strongest issue?	Supplementary question

Student A	I would say being unable to manage my personal finances is my largest issue.	
Interviewer	Anything else?	Prompt
Student A	I think that course material and travel costs have been big issues. I have to pay to come to university on the bus.	
Interviewer	Yes, have you limited the days you come into university?	Supplementary question
Student A	No, not at all, I have always come in when I should do.	
Interviewer	Okay, out of all those six sections, let me remind you what they are: course, institutional, academic, personal, motivational and financial, which one of those six would you say you had the most difficulty with?	Looking back & summing up
Student A	I would say personal, because of the relationship issue.	
Interviewer	Have you had serious motional problems, depression problems then?	Supplementary question
Student A	Yes, with one of the girl friends that I had yes. The biggest two are personal and motivational the fear of failing.	
Interviewer	So that was driving force to help you through your difficulties?	Supplementary question
Student A	Yes	
Interviewer	Did you bury yourself in your work then?	Supplementary question
Student A	I wanted to get this degree.	
Interviewer	Part of your psychi?	Supplementary question
Student A	Yes	
Interviewer	Thanks you very much.	Thanks

Student Analysis

The table below shows a merger of the different responses from all six students according to specific themes. The writing in black is what the students said and is referenced in the left-hand column. The text in red bold italics is supplementary questions asked by the interviewer for clarification. Only relevant comments were placed in the recorded analysis. Only COURSE has been shown, but of course all themes were handled in the same way.

	COURSE
Student A	<p>The course itself would have been, is it too difficult? Is it above me kind of? Can I do this? Erm ..., is the material interesting? The teaching staff on the course as well. I can see that that could come under a different subject area. Will the course lead me into a job that I want? Is the qualification still valid now?</p>
Student B	<p>In my final year I notice a difference, the work became harder and more noticeably than between the first and the second year but I began to enjoy that so I wouldn't say I felt like pulling out because of that. Having said that there was one unit that I decided to defer because I wasn't enjoying it. I didn't think it was taught very well. I did end up deferring that unit and will have to come back next year. So that was the only problem I have had with the course.</p> <p>I chose the course, I didn't rush, I didn't go through Clearing. If I had done it could have affected me. I can understand why some people who start late have difficulty in settling in.</p>
Student C	<p>The course, I was happy with. Moving on to the course now, I didn't have much time because I moved from Sunderland. I think firstly it was the nearest university where I lived really. That was the first .. I just thought Sheffield Hallam. The course, it was really rushed, it was, what can I do, what's similar, the only course that was similar to what I was doing in Sunderland was the Computing one. Which I did look at and I saw points where you could mix it in with different stuff eventually but I found it was different what I read than ..., the subjects were actually different. I think it is an important thing, it was a rushed process. I pleased about it now but ...</p> <p>I think the last year I have struggled because I didn't have a good understand of what involved.</p> <p>I do struggle in things like databases and things, especially programming, which I know lots of people have problems with. I've always not been very good at Maths. I didn't realise that there was Maths at first and how hard that would be. I found that quite hard.</p> <p>I think a couple dropped out and Maths was a major issue in it. I didn't expect it going into it and I read the course guides and things and it wasn't really as I thought it would be with the Maths.</p> <p>I enjoyed multi-media design which I discussed at one point didn't I and I think it annoys you sometimes when you know people, your are trying your hardest and people are doing it and doing what you've done in God knows how many weeks and they do it in a couple of hours. In one particular lesson, (can I mention it? I already have done but) things aren't looked at really from an angle which you aren't expected to look.</p> <p>I enjoyed multi-media design which I discussed at one point didn't I and I think it annoys you sometimes when you know people, your are trying your hardest and people are doing it and doing what you've done in God knows how many weeks and they do it in a couple of hours. In one particular lesson, (can I mention it? I already have done but) things aren't looked at really from an angle which you aren't expected to look.</p>

Student D	<p>I think that my course was appropriate for myself, given my background in statistics and business, because it encapsulated a lot of, a variety of modules that I could use if I chose to go into industry.</p> <p>I found the first year difficult to adapt coming straight from college. I thought that particularly some of the work we were handed involved a lot of group work and given that first years tend not to be so focused on their work, some of the completion of work was quite difficult.</p> <p>At times, I did think it was too easy, given that if you spent the appropriate amount of time on a piece of work you were more than likely to get high marks regardless of the content.</p> <p><i>Did that bore you a bit?</i></p> <p>It did at times because I felt I was just spending time just for the sake of it rather than being interested in the work itself.</p> <p>Generally, the facilities are quite good. At times, when I fine problems was in the library people weren't actually working, it was obvious and the quiet area is not a quiet area at all! There were some other issues with the library. Some of the administration was a bit ... we were never informed when we would get work back when it's been marked. When you have put in a lot of effort to complete it on a date whereas when it's returned we had no idea.</p> <p><i>Wasn't there a due return date?</i></p> <p>For the markings, very rarely as far as I remember.</p>
Student E	<p>At the beginning I think I actually chose a different course, I actually chose Software Engineering. Then, with my predicted grades, I didn't think I would get them so they offered me the Computing course. Now when I got my grades through I did have the grades to do Software Engineering but I'd already preferred the idea of it because I'd been given the prospectus of the Computing course and I decided I wanted to stay on that. It wasn't my first choice at all, I didn't know that the course existed until it was offered to me, it was a brand new course. So I started on that course a little bit late but it was fairly easy to slip into the course, I only missed pretty much the introduction of what the tutors were going to be rather than any learning. So I don't feel I missed too much. I found certain aspects of it, certainly the more enjoyable aspects of it, I found easier. There were some bits that I struggled with that I had to work at.</p> <p><i>Were you in danger of failing any because of it?</i></p> <p>Yes, with my Java I did. I got a referral for that and went back to Liverpool during the holidays I got a tutor at Liverpool University to help me, then when I got back I managed to pass with quite high marks after that. It just needed a bit of one-on-one sort of tutoring. Instead of sharing the issues with the whole class, it took quite a bit of time but I got my head round it in the end.</p> <p><i>That was just that one summer was it where you had to do that?</i></p> <p>No, I slightly struggled again the next year but not to the extent of doing that, two friends back home actually, they were doing Java. They were able to help me through and point me out where I was going wrong and teach me bits and bobs, but it wasn't an actual referral that time.</p> <p><i>Do you think that that might have tripped you up sufficiently to drop out of the course at any stage?</i></p> <p>It could have done, because I am really really not enjoying it. I was struggling and no matter how hard I tried I was still struggling. I was finding it difficult to get a tutor one-on-one, track them down when they were not in lectures and stuff. I was resolving to the internet and books. I was just tried to teach myself out of textbooks and it wasn't happening. I needed someone to sit down, give me examples and explain it.</p>
Student F	<p>No, first year was a bit sketchy. It was as major kind of culture shock coming from A' levels in college, where you are still rather treated like a child, into university. The way that work is given to you is very different, the support you get is very different, much more independent when you come to university. So that was the main difficulty really.</p> <p>To be honest with you, I thought that first and second year were harder than the final year. I don't know whether that's because I went away on placement and they say you mature and you become a different person when you come back. I just that first and second year were very technical, they just didn't seem my kind of</p>

academia. But, then, I tried harder in second year, I got 57% in the first year and 62% in the second year which I was alright with and I was quite happy with 62% and then I did a lot better in the final year.

So, was the course what you expected it to be?

I think they could have mixed up a little bit more kind of the business application side in the first and second year. A lot of stuff, I kind of thought it's the sort of stuff that you're never going to use again. That was in first and second year. Fourth year was very much; you could see how it was applied to industry. A module like Systems Analysis and Design from the way we did it, it just didn't seem like it would ever get done in an organisation. Maybe it is, I really don't know but it just didn't seem as applicable as it could be.

Okay, was that off-putting? Could it have caused you to drop out?

If I'd really struggled and been borderline failing I think maybe yes, but I can't speak for everyone because ...

Appendix B: Semi-Structured Interview Schedule

Introduction

- Thanks very much for agreeing to take part in this interview
- Assurance that what is said is in the strictest confidence
- No reference will be made to you by name
- Anonymous, will be referred to as Student X

Reasons for Research

- Britain has a high dropout rate from undergraduate courses
- Survey to attempt to discover the reasons for high dropout rate

Explanation of procedure

- Series of open-ended questions
- Questions divided up into a number of themes
- Themes – course, institution, academic support, motivation & financial
- You are asked to come up with problems that you might have encountered

Course

- Introduce the theme
- Give examples
 - Rushed choice of university
 - Not what expected etc
- Interact with student and ask supplementary questions
- Thank them when finished theme

Institution

- Introduce the theme
- Give examples
 - Poor study facilities
 - Accommodation problems etc
- Interact with student and ask supplementary questions
- Thank them when finished theme

Academic

- Introduce the theme
- Give examples
 - Fear of examinations
 - Got behind with work etc
- Interact with student and ask supplementary questions
- Thank them when finished theme

Personal

- Introduce the theme
- Give examples
 - Part-time job commitment
 - Lack of self-confidence etc
- Interact with student and ask supplementary questions
- Thank them when finished theme

Motivational

- Introduce the theme
- Give examples
 - Peer group pressure
 - Too many distractions etc
- Interact with student and ask supplementary questions
- Thank them when finished theme

Financial

- Introduce the theme
- Give examples
 - Lack of parental support
 - Unable to manage finances etc
- Interact with student and ask supplementary questions
- Thank them when finished theme

Appendix C: Demographic Choices

1. What is your gender?
 - 1 Male
 - 2 Female

2. What is the highest post-school education of your parents/guardians?
 - 1 None
 - 2 College
 - 3 University
 - 4 Other

3. What is your age group?
 - 1 Under 21
 - 2 21 to 25
 - 3 26 to 30
 - 4 31 to 40
 - 5 Over 40

4. What do you consider your national identity to be?
 - 1 British
 - 2 **English**
 - 3 **Scottish**
 - 4 **Welsh**
 - 5 Northern Irish
 - 6 Irish
 - 7 Other

5. What is your ethnic group?
 - 1 White
 - 2 **White and Black Caribbean**
 - 3 **White and Black African**
 - 4 **Caribbean**
 - 5 **African**
 - 6 **Indian**
 - 7 **Pakistani**
 - 8 **Bangladeshi**
 - 9 **Chinese**
 - 10 **Other**

6. What are your family commitments?
 - 1 None
 - 2 **Dependent children**
 - 3 **Sick partner**
 - 4 **Elderly or sick parent(s)**
 - 5 Other

7. What is your home region?

- 1 **Yorkshire & Humber**
- 2 **East**
- 3 **East Midlands**
- 4 **London**
- 5 **North East**
- 6 **North West**
- 7 **South East**
- 8 **South West**
- 9 **West Midlands**
- 10 **Scotland**
- 11 **Northern Ireland**
- 12 **Ireland**
- 13 **Wales**
- 14 **Other**

8. What is your position within your family?

- 1 First or only child
- 2 **Second** child
- 3 **Third** child
- 4 **Fourth** child
- 5 Greater than fourth child

9. How many brothers and sisters (including half & step) do you have?

- 1 None
- 2 **One**
- 3 **Two**
- 4 **Three**
- 5 More than three

10. Which option best describes the job of your main family wage earner?

- 1 Professional
- 2 **Managerial**
- 3 **Technical**
- 4 **Skilled Non-manual**
- 5 **Partly skilled**
- 6 **Unskilled**

11. Which university do you attend?

(N.B. The other universities that took place in the survey have been anonymised)

- 1 **Sheffield Hallam (SHU)**
- 8 **Other**

12. What sort of accommodation do you live in?

- 1 Rented (furnished)
- 2 **Rented** (unfurnished)
- 3 **Halls of residence**
- 4 **Parental Home**
- 5 **Private** owned
- 6 **Other**

13. Do you have a disability?

- 1 None
- 2 Autistic disorder
- 3 Blind/partially sighted
- 4 ~~Deaf~~hearing impairment
- 5 ~~Learning~~ difficulties
- 6 **Mental** Health difficulties
- 7 **Personal** care support
- 8 **Unseen** (diabetes-epilepsy-asthma)
- 9 **Wheel** chair (mobility difficulties)
- 10 Multiple disorders
- 11 Other

14. What is your marital status?

- 1 Single
- 2 Married
- 3 Separated
- 4 Divorced
- 5 Long-term partner

15. What is your previous educational background?

- 1 School only
- 2 College
- 3 Polytechnic
- 4 University
- 5 Other

16. What is your previous work experience?

- 1 None
- 2 Part-time only
- 3 Under 1 year full-time
- 4 1 to 2 years full-time
- 5 More than 2 years full-time

17. Do you have work experience relevant to your course?

- 1 No
- 2 Yes
- 3 Unsure

18. What year of your course are you currently in?

- 1 First
- 2 **Second**
- 3 **Third**
- 4 **Fourth**
- 5 Over fourth

The first named in each list is the default for that question.

Appendix D: Student Non-Completion Problems

1 Course

- Wrong/rushed course
- Wrong/inadequate skills
- Too difficult
- Not what expected
- Badly designed/taught
- Too long

2 Institutional

- Poor/inadequate study facilities
- Poor/inadequate support facilities
- Accommodation problems
- Administrative problems
- Excluded (tuition, fee debt, academic issues etc)
- Failed or asked to leave

3 Academic

- Lack of study skills (essay writing/note taking)
- Lack of organisational skills
- Concentration problems
- Fear of examinations
- Fear of negative comments from peers/tutors
- Got behind with work
- Failure to submit work
- Lack of attendance at lecture/tutorial

4 Personal

- Difficulties setting in
- Domestic commitments/problems
- Changed circumstances
- Illness/stress
- Relationship/friendship problems
- Lack of self-confidence
- Job commitments
- Non-return/difficulties after gap in studies
- Too many distractions (socialising/recreational)
- Drug/alcohol problems
- Early achievement of goals
- Peer group pressure

5 Financial

- Lack of or inadequate parental support
- Parental/own changed circumstances
- Unable to pay tuition fees
- Travel costs
- Childcare costs etc
- Unable to manage finance

Appendix E: Questionnaire

Page 1 – Demographics



Sheffield Hallam University

STUDENT RETENTION SURVEY

Student University dropout in Britain is currently running at 20%. This **two page** questionnaire is part of an investigation into issues that affect student dropout. By filling in this questionnaire you will be helping to find the major issues that cause dropout and consequently assisting in the improvement of student retention rates. By agreeing to fill in this questionnaire you are giving permission for your answers to be used in this survey. **Your responses are completely anonymous.**

Please answer the following questions about yourself:

1. What is your gender?
Male
2. What is the highest post-school education of your parents/guardians?
None
3. What is your age group?
Under 21
4. What do you consider your national identity to be?
British
5. What is your ethnic group?
White
6. What are your family commitments?
None
7. What is your home region?
Yorkshire & Humber
8. What is your position within your family?
First or only child
9. How many brothers and sisters (including half & step) do you have?
None
10. Which option best describes the job of your main family wage earner?
Professional
11. Which university do you attend?
Sheffield Hallam (SHU)
12. What sort of accommodation do you live in?
Rented (furnished)
13. Do you have a disability?
None
14. What is your marital status?
Single
15. What is your previous educational background?
School only
16. What is your previous work experience?
None
17. Do you have work experience relevant to your course?
No
18. What year of your course are you currently in?
First

Next >>



Sheffield Hallam University

STUDENT RETENTION SURVEY

Please indicate your level of agreement with the following statements using the scale below:

1	2	3	4	5
Strongly Disagree	Tend to Disagree	Undecided	Tend to Agree	Strongly Agree

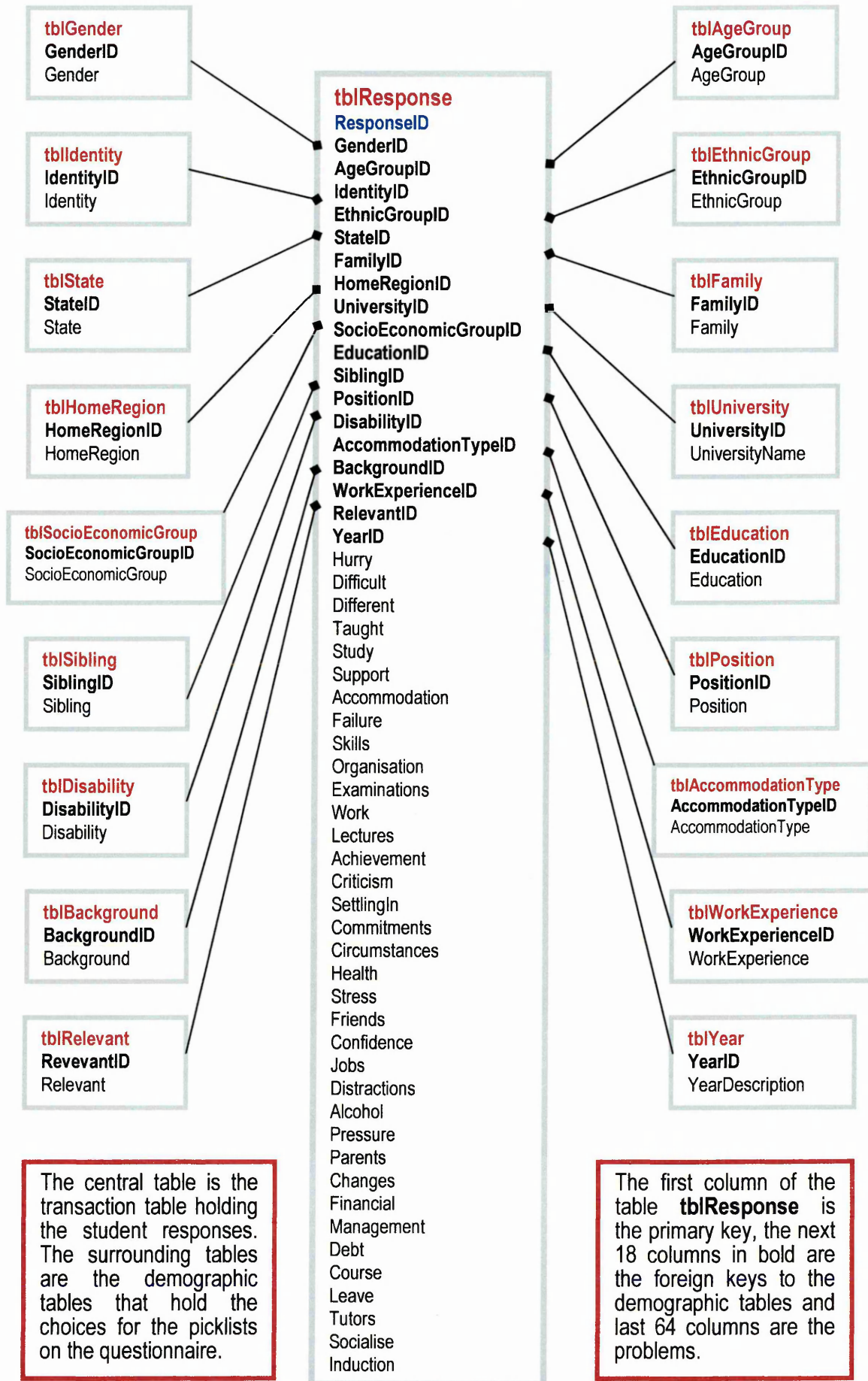
Please click on a number from 1 to 5 for each of the following statements:

- | | |
|--|---|
| 1. I chose my course in a hurry | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 2. The course has been too difficult | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 3. The course is different from what I expected | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 4. The course has been well taught | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 5. The study facilities are good (library, computers etc) | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 6. The support facilities are good | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 7. I have had no difficulties with student accommodation | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 8. I have had more failures or referrals than most students | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 9. At the start of my course I lacked basic skills such as essay writing & note taking | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 10. My organisational skills are good | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 11. I do not perform to my best in examinations | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 12. I have always kept up with my work | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 13. I have missed more lectures and tutorials than most other students | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 14. I am satisfied with my level of achievement | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 15. I dislike criticism from tutors and/or peers | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 16. I had difficulty settling in when I came to University | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 17. I have a number of commitments outside University | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 18. My personal circumstances have changed whilst being at University | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 19. I have had health problems that have caused me to take time off | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 20. I have never suffered from stress | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 21. I have had a number of friendship/relationship problems | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 22. I lack self confidence | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 23. My part-time job conflicts with my studying | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 24. There are too many distractions that affect my ability to study | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 25. I have had issues with drugs and/or alcohol | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 26. I give in to peer-group pressure | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 27. I have always had parental financial support | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 28. My family's financial circumstances have changed during my time at University | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 29. I have many financial commitments | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 30. I am bad at managing my finances | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 31. I have a large debt apart from my student loan | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 32. I have sometimes found the course stressful | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 33. I have considered changing/leaving at some stage | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 34. My tutors give me support when I need it | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 35. I have found it easy to make friends at University | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |
| 36. The induction programme helped me to feel more comfortable at University | <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 |

1	2	3	4	5
Strongly Disagree	Tend to Disagree	Undecided	Tend to Agree	Strongly Agree

Submit

Appendix F: Input Database Schema



Appendix G: Variables and Types

The Input Database is made up of two types of table, the demographic tables (Gender, Age Group etc) and the response table. These are set out below. **PK** indicates the Primary Key to the table and **FK** indicates a Foreign Key.

Demographic Tables			Response Table	
	tblGender		tblResponse	
PK	GenderID	smallint	PK ResponseID	smallint
	GenderName	varchar(7)	FK GenderID	smallint
	tblAgeGroup		FK AgeGroupID	smallint
PK	AgeGroupID	smallint	FK IdentityID	smallint
	AgeGroup	varchar(9)	FK EthnicGroupID	smallint
	tblIdentity		FK StateID	smallint
PK	IdentityID	smallint	FK FamilyID	smallint
	Identity	varchar(16)	FK HomeRegionID	smallint
	tblEthnicGroup		FK UniversityID	smallint
PK	EthnicGroupID	smallint	FK SocioEconomicGroupID	smallint
	EthnicGroup	varchar(26)	FK EducationID	smallint
	tblState		FK SiblingID	smallint
PK	StateID	smallint	FK PositionID	smallint
	State	varchar(21)	FK DisabilityID	smallint
	tblFamily		FK AccommodationTypeID	smallint
PK	FamilyID	smallint	FK BackgroundID	smallint
	Family	varchar(26)	FK WorkExperienceID	smallint
	tblHomeRegion		FK RelevantID	smallint
PK	HomeRegionID	smallint	FK YearID	smallint
	HomeRegion	varchar(21)	Hurry	tinyint
	tblUniversity		Difficult	tinyint
PK	UniversityID	smallint	Different	tinyint
	UniversityName	varchar(26)	Taught	tinyint
	tblSocioEconomicGroup		Study	tinyint
PK	SocioEconomicGroupID	smallint	Support	tinyint
	SocioEconomicGroup	varchar(41)	Accommodation	tinyint
	tblEducation		Failure	tinyint
PK	EducationID	smallint	Skills	tinyint
	Education	varchar(11)	Organisation	tinyint
	tblSibling		Examinations	tinyint
PK	SiblingID	smallint	Work	tinyint
	Sibling	varchar(16)	Lectures	tinyint
	tblPosition		Achievement	tinyint
PK	PositionID	smallint	Criticism	tinyint
	Position	varchar(26)	SettlingIn	tinyint
	tblDisability		Commitments	tinyint
PK	DisabilityID	smallint	Circumstances	tinyint
	Disability	varchar(41)	Health	tinyint
	tblAccommodationType		Stress	tinyint
PK	AccommodationTypeID	smallint	Friends	tinyint
	AccommodationType	varchar(26)	Confidence	tinyint
	tblBackground		Jobs	tinyint
PK	BackgroundID	smallint	Distractions	tinyint
	Background	varchar(13)	Alcohol	tinyint
	tblWorkExperience		Pressure	tinyint
PK	WorkExperienceID	smallint	Parents	tinyint
	WorkExperience	varchar(31)	Changes	tinyint
	tblRelevant		Financial	tinyint
PK	RelevantID	smallint	Management	tinyint
	Relevant	varchar(11)	Debt	tinyint
	tblYear		Course	tinyint
PK	YearID	smallint	Leave	tinyint
	YearDescription	varchar(16)	Tutors	tinyint
			Socialise	tinyint
			Induction	tinyint

Appendix H: Conversion Criteria

Statement Comparison

OLD STATEMENT

1. I chose my course in a hurry
2. The course has been too difficult
3. The course is different from what I expected
4. The course has been well taught
5. The study facilities are good (library, computers etc)
6. The support facilities are good
7. I have had no difficulties with student accommodation
8. I have had more failures or referrals than most students
9. At the start of my course I lacked basic skills such as essay writing & note taking
10. My organisational skills are good
11. I do not perform to my best in examinations
12. I have always kept up with my work
13. I have missed more lectures and tutorials than most other students
14. I am satisfied with my level of achievement
15. I dislike criticism from tutors and/or peers
16. I had difficulty settling in when I came to University
17. I have a number of commitments outside University
18. My personal circumstances have changed whilst being at University
19. I have had health problems that have caused me to take time off
20. I have never suffered from stress
21. I have had a number of friendship/relationship problems
22. I lack self confidence
23. My part-time job conflicts with my studying
24. There are too many distractions that affect my ability to study
25. I have had issues with drugs and/or alcohol
26. I give in to peer-group pressure
27. I have always had parental financial support
28. My family's financial circumstances have changed during my time at University
29. I have many financial commitments
30. I am bad at managing my finances
31. I have a large debt apart from my student loan
32. I have sometimes found the course stressful
33. I have considered changing/leaving at some stage
34. My tutors give me support when I need it
35. I have found it easy to make friends at University
36. The induction programme helped me to feel more comfortable at University

NEW STATEMENT

1. Hurried choice of course
2. Difficult course
3. Course different from expectation
4. Taught badly
5. Facilities unsatisfactory
6. Support unsatisfactory
7. Accommodation difficulties
8. Failures more than most
9. Skill deficiency at start of course
10. Organisation skills poor
11. Examination performance below expectations
12. Behind with work
13. Missed more lectures and tutorials than most
14. Achievement below expectation
15. Criticism disliked
16. Settling in was difficult
17. Outside commitments high
18. Personal circumstances changed
19. Time off with health problems
20. Stress problems
21. Friendship or relationship problems
22. Lack self confidence
23. Part-time job conflicts with studies
24. Too many distractions that affect ability to study
25. Issues with drugs and/or alcohol
26. Give in to peer-group pressure
27. Parental financial support lacking
28. Family's financial circumstances changed
29. Financial commitments high
30. Bad at managing finances
31. Large debt apart from my student loan
32. Have sometimes found course stressful
33. Considered changing/leaving at some stage
34. Tutor support sometimes lacking
35. Found it difficult to make friends at University
36. The induction didn't help to feel more comfortable at University

Conversion Notes

The left-hand column above contains the original problems as shown on the questionnaire. The right-hand column contains the reworded statements. Most of them are more concisely worded but those in red were rewritten so that the 'problem' became the 'Yes' response. For instance statement 4 has been reworded from 'The course has been well taught' where a response of 'Yes' would indicate no problem, to 'Taught badly' where a 'Yes' response would indicate a problem with the teaching.

All responses of 1 and 2 are put together (Strongly disagree and Tend to disagree) and responses of 4 and 5 are put together (Tend to agree and Strongly agree). In ALL cases a response of 3 (Undecided) has been put with the 'No' response. This may well have limited the research findings, but the author believes that false negatives are preferable to potentially false positives.

Appendix I: Data Mining Dataset

Response

Gender
AgeGroup
Identity
EthnicGroup
State
Family
HomeRegion
UniversityName
SocioEconomicGroup
Education
Sibbling
Position
Disability
AccommodationType
Background
WorkExperience
Relevant
YearDescription
Hurry
Difficult
Different
Taught
Study
Support
Accommodation
Failure
Skills
Organisation
Examinations
Work
Lectures
Achievement
Criticism
SettlingIn
Commitments
Circumstances
Health
Stress
Friends
Confidence
Jobs
Distractions
Alcohol
Pressure
Parents
Changes
Financial
Management
Debt
Course
Leave
Tutors
Socialise
Induction

Response Discussion

All the questions of the original database have been merged into one flat table. The first eighteen columns from **Gender** to **YearDescription** are from Page 1 of the input form and the remaining thirty-six columns from **Hurry** to **Induction** are the problem statements from Page 2. See Appendix E for the input form.

Unlike the original database as shown in Appendix F there is only one table. The choices of each the columns in the single table **Response** from **Gender** to **YearDescription** are as shown in Appendix C.

The remaining thirty-six columns are Yes/No responses to the problems posed in Appendix D. These are recorded as Y or N. However, the problem statements were reworded to ensure that the YES (Y) response reflected the problem. (See Section 9.3(c))

Appendix J contains a table showing the Data Mining statement alongside the database variables in the column opposite.

A comparison of the original statements alongside the revised statements can be found in Appendix H above.

Appendix J: Problem to Variable Lookup Table

DATA MINING PROBLEMS	DATABASE VARIABLES
1. Hurried choice of course	Hurry
2. Difficult course	Difficult
3. Course different from expectation	Different
4. Taught Badly	Taught
5. Facilities unsatisfactory	Study
6. Support unsatisfactory	Support
7. Accommodation difficulties	Accommodation
8. Failures more than most	Failure
9. Skill deficiency at start of course	Skills
10. Organisation skills poor	Organisation
11. Examination performance below expectations	Examination
12. Behind with work	Work
13. Missed more lectures and tutorials than most	Lectures
14. Achievement below expectation	Achievement
15. Criticism disliked	Criticism
16. Settling in was difficult	SettlingIn
17. Outside commitments high	Commitments
18. Personal circumstances changed	Circumstances
19. Time off with health problems	Health
20. Stress problems	Stress
21. Friendship or relationship problems	Friends
22. Lack self confidence	Confidence
23. Part-time job conflicts with studies	Jobs
24. Too many distractions that affect ability to study	Distractions
25. Issues with drugs and/or alcohol	Alcohol
26. Give in to peer-group pressure	Pressure
27. Parental financial support lacking	Parents
28. Family's financial circumstances changed	Changes
29. Financial commitments high	Financial
30. Bad at managing finances	Management
31. Large debt apart from my student loan	Debt
32. Have sometimes found course stressful	Course
33. Considered changing/leaving at some stage	Leave
34. Tutor support sometimes lacking	Tutors
35. Found it difficult to make friends at University	Socialise
36. The induction didn't help to feel more comfortable at University	Induction

Appendix K: Post Enquiry Interviews

Appendix K.1 Introduction

Now that the major issues have been isolated it is necessary for them to be verified by interviewing a number of students that have had significant problems in the current academic year 2005-6. All the students chosen for interview had either left the university (dropped out of their courses) or had significant referrals at the end of the year that made it unlikely that they will complete the year. The students were located from the data prepared for the Assessment Boards in June 2006. Only the most serious cases were considered. It was decided to conduct the interviews by telephone using a fully-structured interview schedule that was designed to take only ten minutes. This meant that the students only needed to be contacted once, and if they agreed the interview could be conducted immediately. Ten minutes was specified so that the task wasn't considered to be too onerous by the interviewee. The interview schedule can be found in Appendix L. These interviews were designed to achieve Objective 4, *"Validate the results of the quantitative research within the ITPA by means of structured interviews"*. (See Section 2.2) In addition, all interviews were conducted between the 7th and 14th of July 2006.

Appendix K.2 The Interviews

There was a significant problem locating students. Because of the problem of confidentiality it was decided not to ring the students on their home number (generally the parental home) as this might compromise the student. Instead students were contacted on the last known mobile telephone number. Many of these contact telephone numbers gave the unobtainable signal, suggesting that the number had changed. Even when the number responded there was often no reply. In these cases the number was rung on three separate occasions before abandoning the efforts to contact the student. For every student that was successfully contacted there were another three that could not be contacted. Telephone interviews were chosen since postal responses or face to face contacts was thought to be too time consuming and unlikely to be achieved in many cases. Three students were selected for interview from each year of study.

Once a student had been successfully contacted they were told of the purpose of the survey and the confidentiality of their responses. All interviewees agreed to take part in the survey. The interview schedules were fully structured allowing little room for individuality. The purpose of this was to ensure conformity of response to ease the task of analysis. The students were offered the opportunity to give reasons for their lack of progress. The intention here was to allow students to say in their own words what the main issues were that had caused their lack of progress. This response was asked for before the findings of the study were introduced so as not to bias their responses. The major demographic and problem issues from the research findings were then put to the student for comment. A précis of the interviews can be found in Appendix M.

It was decided to ask an additional demographic question about whether they had been enrolled on a university course previously. This turned out to be a fruitless exercise since only one student had. All the problem issues were asked, even those that had only been found by one of the techniques. On balance, however, it was thought that evidence from Rules 1 and 2 had a stronger base as there was a strong correlation between them (see Section 9.5.2).

Appendix K.3 Analysis of the Interviews

Appendix K.3.1 Initial Impressions

The analysis starts by looking at the data generally and getting an overall impression. Nine students were interviewed, three first years, three second years and three from final year (one third and two fourth). The fourth year students had taken a year out from studying, Student 9 as a placement year (48 weeks) relevant to the course and Student 8 in an unrelated activity. All students had less than a year's full-time work experience, most had only part-time experience. Only one student (Student 8) had any work experience relevant to the course. All students admitted to getting behind with their work and all but one student believed that there had been too many distractions that affected his/her ability to study. Eight out of the nine also admitted to having a high level of outside commitments.

Twenty-two issues were investigated, but only eighteen of these were considered since the last four were felt to be of lesser significance (see 10.2 above). The responses in red in Appendix M show the students in the at-risk group; that is, at-risk of changing or leaving their course. Of the remaining eighteen issues, no student scored less than nine out of eighteen. That is, the students were all in the at-risk group for at least half of the issues. Student 5 scored thirteen out of eighteen.

Rating the issues according to the number of students falling into them was as follows:

PROBLEM	Tally
Less than 1 year work experience or only part-time	9
Got behind with work	9
No work experience relevant to course	8
Too many distractions that affected their ability to study	8
Outside commitments high	8
Have had stress problems	7
First or second year	6
Non-local	6
Younger sibling	6
Have sometimes found the course stressful	5
Personal circumstances changed	5
Bad at managing finances	4
Course different from expectations	4
Under 21	3
Suffer from lack of self-confidence	3
From a large family (3 or more children)	2
Examination performance has been below expectations	2
Halls of residence	1

Appendix K.3.2 A Closer Look

The students were first asked what they considered to be the issues that had been the major contributor to their failure to complete the year. These generally fitted with their responses to the problem issues. Many were personal problems at home, such as the death of a close relative or relationship issues such as girl friend becoming pregnant (Student 4). Student 9 said that his father had been in hospital for a significant amount of time and he had been left to look after the family, (five younger brothers and sisters) whilst having to sustain a part-time job. He stated that he had not had enough time to devote to the course and to revision for the final examinations. Student 3 stated that the lack of feedback from lecturers due to the lecturer dispute had been a major contributory factor. Student 5 admitted to serious depression that had stopped him attending university altogether. Student 3 stated that he had not adapted to the course at all well and that it was significantly different from what he had expected. He had

done an ICT related course at a Further Education College prior to coming to university and expected his degree course to be similar. He was shocked with the amount of programming that he had had to do. However, some of the students admitted to having had difficulties previously and that they were carrying forward modules from previous years. They cited this as a contributory factor to their current difficulties. "I was carrying a module from 2nd year". (Student 8)

Appendix K.3.3 Looking at the Major Issues

From the students surveyed it appears that the lack of work experience, in particular work relevant to their course was a major common factor. Indeed this was the case for almost all the students. The combination of too many distractions and high outside commitments seem to have a strong contribution to them getting behind with their work. Stress and depression (Student 5) rank highly with these students. Six out of the nine were younger siblings and more than half admitted that their personal circumstances had significantly changed during their time at university. However, five of them had lived in the parental home whilst studying at the university, although one had travelled daily from Leeds. Only one had lived in Halls of Residence, though three of them had done so in their first year.

Appendix K.4 Summary

Overall the interviews appear to back up the main findings of this study. Clearly this is only a sample reflective view, but seems to suggest that the findings are sensible. The author would have liked to have sampled a greater number of students but concluded that on balance nothing further would have been found should a larger sample have been taken. This was in part influenced by the degree of difficulty encountered in reaching these students. However, it was felt that this sort of exercise was worthwhile repeating on a regular basis as a means of follow-up and student care.

It is recommended that an interview schedule should be set up to be trialled by the ITPA for use in the next academic year. It is felt that the exercise should be started early in the academic year, say after the first six to eight weeks to ensure that students most at risk were contacted prior to them leaving. The basis for

interview should be lack of attendance, since it is unlikely that any assessment will have taken place at this early stage. The use of 'lack of attendance' rather than actual drop-out was justified as it was assumed to be a clear sign that the student might be considering dropping out.

Once collected, the data should then be passed to the author for analysis. The results of this data analysis should then be passed to FARG for consideration, revision and action.

Appendix L: Structured Student Interviews

Hello, I'm Keith Burley from Sheffield Hallam University Faculty of ACES. I believe that you have had problems this year that have resulted in non-completion of the year.

Student response

I am conducting some research into student non-completion in an attempt to give more support to vulnerable students. I wonder if I could ask you a few questions. It won't take very long?

Student response

What were the main issues that you encountered that have caused your non-completion this year?

Student response

Thank you very much for that. Could I now ask you question about yourself please?

Student response

Demographic	Student Response
What is your age?	
What is your year of study?	
Have you been enrolled on a university course previously?	
What town or city do you come from?	
How many brothers and sisters do you have?	
What is your position within these siblings?	
What sort of accommodation did you live in whilst at university?	
Do you have any work experience? If so, what sort?	
Do you have any work experience relevant to the course that you are/were taking?	

Could I now ask you some specific questions please, about possible problems you might have had?

Student Response

Problem	Student Response
Did you sometimes find the course stressful?	
Have you ever suffered from stress?	
Were there too many distractions that affected your ability to study?	
Did you perform badly in examinations?	
Are you bad at managing your finances?	
Was the course different from your expectations?	
Do you suffer from lack of self-confidence?	
Are your outside interests high?	
Did you get behind with your work?	
Have your personal circumstances changed whilst you have been at university?	
Did the induction course help you to feel more comfortable?	
Do you dislike criticism?	
Do you feel that you have been taught badly?	
Is there anything else you would like to say?	

Thank you very much for agreeing to take part in this study.

Keith M Burley

5th July 2006

Appendix M: Analysis of Post-Survey Student Interviews

PROBLEM	Stud 1	Stud 2	Stud 3	Stud 4	Stud 5	Stud 6	Stud 7	Stud 8	Stud 9
Age	20	18	19	24	23	23	22	22	22
Year	1st	1st	1st	2nd	2nd	2nd	3rd	4th	4th
Town / City of origin	Sheffield	York	Leeds	Wigan	All over	Angola	Sheffield	Doncaster	Sheffield
No of brothers and sisters	1	0	1	2	3	0	1	1	9
Position in family	1st	1st	2nd	2nd	3rd	1st	2nd	2nd	3rd
Accommodation whilst at university	Parental Home	Halls	Parental Home	House (furnished)	House (furnished)	House (furnished)	Parental Home	Parental Home	Parental Home
Work experience	< 1 year	P/T only	P/T only	P/T only	P/T only	P/T only	< 1 year	< 1 year	< 1 year
No work experience relevant to course	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Found course stressful	No	No	Yes	No	Yes	No	Yes	Yes	Yes
Suffered from stress	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes
Too many distractions	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Perform badly in examinations	No	No	Yes	No	No	No	No	Yes	No
Bad at managing finances	Yes	No	No	Yes	Yes	No	No	No	Yes
Course different from expectations	Yes	No	Yes	No	No	Yes	No	No	Yes
Suffer from lack of self-confidence	No	No	No	No	Yes	Yes	No	No	Yes
Outside interests high	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Got behind with work *	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Personal circumstances changed *	Yes	No	No	Yes	Yes	No	Yes	No	Yes
Induction didn't help to settle in *	No	No	Yes	No	Yes	No	Yes	No	No
Dislike criticism *	No	No	Yes	No	No	No	No	No	No
Badly taught *	No	No	No	No	Yes	Yes	No	No	Yes
Been to university before *	No	No	No	No	Yes	No	No	No	No
Other issues that affected you	Personal problems at home	Lack of feedback due to lecturers action	Not adapted to course well, different from what expected	Death in family, girl friend pregnant	Depression, stopped going to university	Holiday in January, illness, felt low	Too much going on outside university	The amount of time for assignments etc, carrying a module from 2nd year	Family difficulties, not enough time for revision
TALLY	11	9	12	10	13	10	9	9	13
POSSIBLE	18	18	18	18	18	18	18	18	18